VR-lähiliikenteen matkustajamäärien estimointi sekä matkan pituuksien mallintaminen automaattisilla matkustajalaskentalaitteilla kerättyjen näytteiden perusteella

Noora Nikula

Helsingin Yliopisto

Matemaattis-luonnontieteellinen tiedekunta

Matematiikka

Pro gradu-tutkielma

Toukokuu 2013

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Matematiikan ja tilastotieteen laitos

Tekijä — Författare — Author
Noora Nikula

Työn nimi — Arbetets titel — Title
VR-lähiliikenteen matkustajamäärien estimointi sekä matkan pituuksien mallintaminen automaattisilla matkustajalaskentalaitteilla kerättyjen näytteiden perusteella

Oppiaine — Läroämne — Subject
Matematiikka

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Toukokuu 2013	62 s.

Tiivistelmä — Referat — Abstract

Tutkielmassa tarkastellaan VR:n lähiliikenteen matkustajamäärien tilastollista estimointia sekä yksittäisen matkustajan matkan pituuden mallintamista automaattisten matkustajalaskentalaitteiden keräämän aineiston perusteella. Tutkielma on osa HSL:n ja VR:n yhteistä lähijunaliikenteen matkustajamäärätutkimuksen uudistamishanketta.

Keväällä 2013 lähijunaliikenteen junista noin 25 % sisälsivät matkustajalaskentalaitteet. Tutkielmassa käsitellään laskentalaitteiden keräämän aineiston matkustajamäärien estimointiin asettamia haasteita, kuten aineiston vinoumaa, erävastauskatoa sekä kehikkovirheitä. Lisäksi pohditaan painotusmenetelmien sekä imputoinnin sopivuutta vastauskadon oikaisumenetelmänä. Tutkielman tavoitteena on kehittää täysin uusi tiedonjalostusprosessi, jonka pohjalta lähijunaliikenteen matkustajamäärätilastot toteutetaan kuukausittain.

Lähijunaliikenteen lähtöpopulaatio on jakautunut aikataulujen määräämiin homogeenisiin ryhmiin. Lyhyellä aikavälillä ei matkustajien käyttäytyminen muutu merkittävästi, jolloin aikataulunmukaisten lähtöjen realisointeja voidaan pitää melko samoinjakautuneina. Riippuen linjasta joka kuukausi esiintyy enemmän tai vähemmän aikataulunmukaisia lähtöjä, joista ei saada yhtään mittausta. Imputoinnin uskottiin olevan painotusmenetelmää joustavampi vaihtoehto reagoida vastauskatoon, koska siinä imputoitavien arvojen luovuttajaa voidaan etsiä kuukaudelta, jota voidaan pitää matkustajien käyttäytymisen kannalta samankaltaisena.

Laskentalaitteet eivät merkitse matkustajia, mistä syystä yksittäisen matkustajan kulkeman matkan pituus ei ole aineistosta suoraan havaittavissa. Jokaiselle lähdölle voidaan muodostaa yksittäisten matkojen jakauman toteuttama lineaarinen systeemi, jolle ei kuitenkaan usein löydy yksikäsitteistä ratkaisua. Kuitenkin riippumattomien ja samoinjakautuneiden näytteiden perusteella voidaan löytää matkan pituuden todennäköisyysjakauma ehdolla matkustajan lähtöasema.

Tutkielmassa käytetään Bayesilaista malliestimointia, jossa mallin piilomuuttuja on yksittäisten matkojen jakauma. Posteriorijakaumaa tutkitaan Markovin ketjun Monte Carlo-menetelmillä. Idea on lähteä konstruoimaan Markovin ketjua, jonka tasapainojakauma on haluttu posteriorijakauma. Tarkastelua syvennetään todistamalla suurten lukujen laki sekä keskeisen raja-arvolauseen erikoistapaus Markovin ketjulle seuraamalla Esa Nummelinin artikkelia *MC's for MCMC'ists*.

Malliestimoinnin lähteenä käytetään myöhemmin ilmestyvää *A local train problem*-artikkelia (Gasbarra D. et al.), jossa malliestimointia tarkastellaan perusteellisemmin.

Avainsanat — Nyckelord — Keywords
vastauskato, painotusmenetelmät, imputointi, Markovin ketju, MCMC-menetelmät

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Alkusanat

Tämä Pro gradu-tutkielma on tehty osana HSL:n ja VR:n yhteistä lähijunaliikenteen matkustajamäärätutkimuksen uudistamishanketta, jonka tarkoituksena oli aloittaa hyödyntää automaattisten matkustajalaskentalaitteiden tuottamaa aineistoa matkustajamäärien tilastoinnissa.

Ensiksi haluan kiittää HSL:llää ja VR:ää mahdollisuudesta tämän tutkielman tekemiseen. Erityisesti haluan kiittää projektiin osallistuneita työryhmän jäseniä, joiden kanssa on ollut mukava työskennellä ja joiden työpanoksen sekä tietotaidon avulla projekti saatiin läpivietyä. Lisäksi haluan kiittää koko HSL:n operatiiviset tutkimukset ryhmän jäseniä mukavan työilmapiirin luomisesta.

Kiitän yliopistonlehtoria Dario Gasbarraa ohjauksesta ja neuvoista, jotka ovat tutkielman kannalta olleet korvaamattomia.

Helsingissä 17.5.2013

_____

Noora Nikula

Sisältö

1 Johdanto **1**

1.1 Tutkielman tausta 1

1.2 Tutkielman tavoitteet 2

2 APC-laitteet lähijunaliikenteessä **4**

2.1 Laskentalaitteet ja aineiston tietokanta 4

2.2 Tutkimusongelmia 6

2.2.1 Ei-todennäköisyysotantaan perustuva otos 6

2.2.2 Erävastauskato 7

2.2.3 Kehikkovirheet 8

2.3 APC-mittausten vertaaminen käsinlaskentoihin 9

3 Vastauskadon oikaisumenetelmiä **10**

3.1 Painotusmenetelmät 10

3.1.1 Peruspainot 11

3.1.2 Jälkiosittaminen 12

3.2 Imputointi 13

4 Matkustajamäärätutkimus lähijunaliikenteessä **17**

4.1 Vastauskadon oikaisumenetelmän sekä raportointisovelluksen valinta 17

4.1.1 DavisWebin painotusmenetelmät 17

4.1.2 Painotusmenetelmät vai imputointi? 18

4.2 Tilastollinen editointi ja imputointi 21

4.2.1 Tilastollinen editointi DavisWebissä 21

4.2.2 Tilastollinen editointi SAS-ohjelmistossa 22

4.2.3 Imputointimenetelmät 24

4.3 Raportointi 27

5 Markovin teoriaa **29**

5.1 Stokastiset prosessit 29

5.2 Satunnaismuuttuja ja todennäköisyysavaruus 29

5.3 Markovin ketju 31

5.3.1 Markov-ominaisuus 31

5.3.2 Siirtymäydin 32

5.4 Suurten lukujen laki Markovin ketjulle 33

5.4.1 Uusiutuminen . . . 34
5.4.2 Potentiaalifunktio . . . 37
5.4.3 Uusiutumisajan odotusarvon äärellisyys . . . 38
5.4.4 Invariantti jakauma . . . 40
5.4.5 Rekurssiivisuus . . . 41
5.4.6 SLL:n todistus . . . 43
5.5 Keskeinen raja-arvolause Markovin ketjulle . . . 44
5.5.1 Ergodisuus . . . 44
5.5.2 KRL:n todistus . . . 49
5.6 Markovin piilomalli . . . 52
5.7 Markovin ketjun Monte Carlo-menetelmä . . . 53
5.7.1 Metropolis-Hastings-algoritmi . . . 53

6 Markovin piilomalli lähijunaliikenteessä **55**
6.1 Mallin parametrit . . . 55
6.2 Malli . . . 57

7 Lopuksi **59**

Lähdeluettelo **61**

1 Johdanto

1.1 Tutkielman tausta

Lähijunaliikenteen automaattinen matkustajalaskenta on Helsingin seudun liikenne -kuntayhtymän (HSL) ja VR-Yhtymä Oy:n (VR) yhteinen Business Intelligence-projekti, jonka päämäärä on nykyaikaistaa Helsingin seudun lähijunaliikenteen matkustajamäärätutkimus. HSL, VR ja Pääkaupunkiseudun Junakalusto Oy (JKOY) ovat investoineet juniin sijoitettaviin automaattisiin matkustajalaskentalaitteisiin (*Automatic Passenger Counter, APC*), jotka laskevat nousevat ja poistuvat matkustajat jokaiselta asemalta. Ensimmäinen tutkimuskysymys käsittelee lähijunaliikenteen aggregaattitason matkustajamäärien tilastollista estimointia APC-laitteiden keräämien näytteiden perusteella. Koska matkustajamääräestimaateilla on keskeinen merkitys joukkoliikenteen kehittämisessä, muodostavat tiedonhallinta ja raportointi yhden projektin onnistumisen mittarin, eikä niitä siksi tulla täysin sivuuttamaan. Toisessa tutkimuskysymyksessä näytetään, että käyttämällä ainoastaan laskentalaitteiden tuottaamaa aineistoa, yksittäisen matkan todennäköisyysjakauma ehdolla matkustajan lähtöasema on tunnistettavissa.

Matkustajamäärätutkimusta on tehty Helsingin seudun lähijunaliikenteessä ainakin vuodesta 1988 asti. Aineistoa on kerätty käsinlaskennoin, joita on suoritettu vuosina 1988, 1989 ja 1990 lokakuussa sekä vuodesta 1991 eteenpäin lisäksi joko maalis- tai huhtikuussa riippuen pääsiäisen ajakohdasta. [17, 2012.] Jokakeväinen ja -syksyinen otos on sisältänyt yhden arki-, lauantai ja sunnuntaipäivän kaikki suunnitellut lähdöt, joiden nousijat ja poistujat on laskettu asemittain. Asemalaskentaa ovat suorittaneet pääasiassa urheiluseurat, partiolaiset tai muut vastaavat tahot. Myöhäisillan asemalaskennasta sekä paikkauslaskennasta on vastannut junahenkilökunta. Käsinlaskennalla poimitun otoksen matkustajamäärät on yleistetty koko vuoden kokonaismatkustajamääräestimaateiksi kiinteiden vakiokertoimien avulla. Viimeinen käsinlaskenta suoritettiin keväällä 2012 ja APC-laitteiden keräämiin näytteisiin pohjautuvaan kuukausiraportointiin oli aikomus siirtyä vuoden 2013 tammikuusta lähtien. Kappaleessa (2.2) lukija perehdytetään tarkemmin APC-laitteiden keräämän aineiston ongelmakohtiin lähijunaliikenteessä. Suurin yksittäinen ongelma on aineiston valikoituneisuus. Tämän oikaisemiseksi kappaleessa (3) esitetään kaksi vastauskadon oikaisumenetelmää: painotusmenetelmät ja imputointi. Kappaleessa (4) pohditaan edellisten menetelmien soveltuvuutta lähijunaliikenteeseen.

Yksittäisen matkustajan käyttäytymistä on tutkittu erilaisten kyselytutkimusten perusteella jo vuodesta 1960 lähtien. Uusin syksyllä 2012 tehdyn liikkumistutkimuksen tavoite oli saada luotettava kuva siitä, kuinka Helsingin seudun asukkailla on tapana liikkua

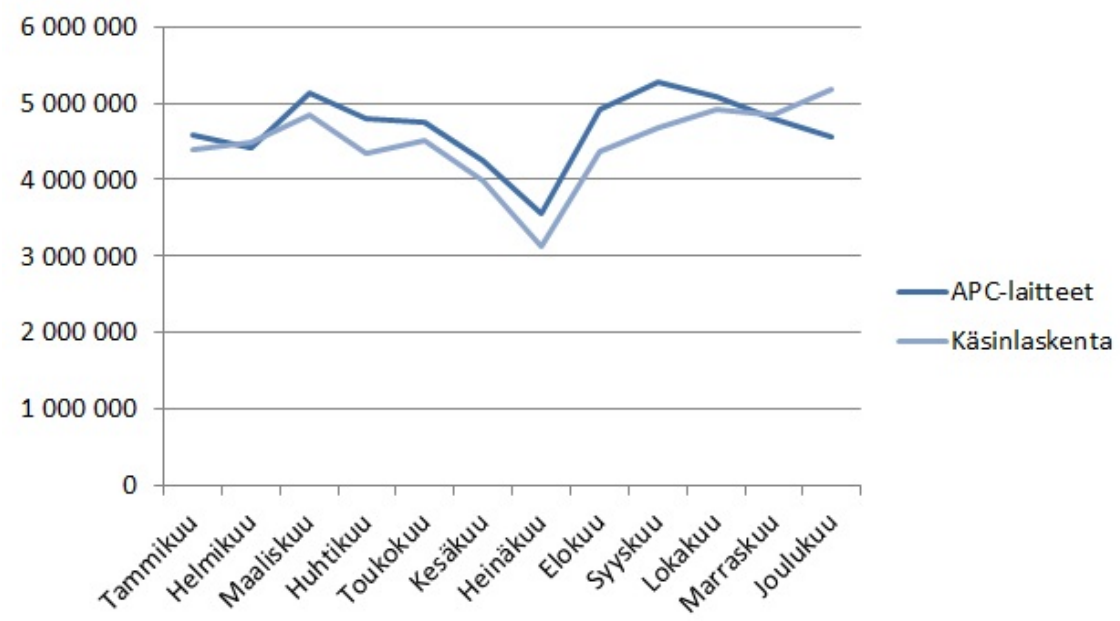
ja millaiset asiat liikkumiseen vaikuttavat. Tutkimus on osa Helsingin seudun liikennejärjestelmäsuunnittelua. [9, 2012.] APC-laitteiden avulla ei voida mitata matkustajien liikennetottumuksiin vaikuttavia tekijöitä, eikä niiden avulla saada tietoa koko matkaketjusta, mutta ne tarjoavat nopean ja edullisen tavan tutkia matkustajan mikrotason käyttäytymistä. Kappaleessa (5) lukija tutustutetaan Markovin ketjuihin ja Markovin piilomalleihin, jonka jälkeen kappaleessa (5.7) esitetään Markovin ketjun Monte Carlo-menetelmä matemaattisen mallintamisen apukeinona. Lopulta kappaleessa (6) teoriaa sovelletaan lähijunaliikenteen yksittäisen matkan pituuden mallintamiseen.

Matkustajamäärien merkitys joukkoliikenteen kehittämisessä sekä linjasto- ja aikataulusuunnittelussa on merkittävä. Lähdön matkustajamäärä sekä maksimikuorma ovat erinomaisia vuoron kysynnän mittareita, jotka tarjoavat lähtökohdan palvelutason, kuten vuorojen tiheyden sekä kaluston käytön, suunnitteluun. Junaliikenteessä matkustajamäärillä on myös taloudellinen merkitys; ne määräävät yhdessä $3-5$ vuoden välein suoritettavan lippulajitutkimuksen kanssa HSL:n VR:lle maksamia lipputulomenetyskorvauksia YTV-alueen ulkopuolella. Kokonaisuudessaan joukkoliikenteen matkustajamäärä on yksi liikennelaitoksen menestyksen mittari, sillä se on keskeinen avaintunnusluku joukkoliikenteen kehityksen seurannassa.

1.2 Tutkielman tavoitteet

Siirtyminen käsinlaskennoista automaattisiin laskentalaitteisiin oli tuottanut positiivisia kokemuksia Helsingin raitiovaunuliikenteessä. Käsinlaskentoja jatkettiin vertailun vuoksi vielä laitteiden käyttöönoton jälkeenkin. Kuvasta (1) nähdään, että manuaalisesti kerätyn aineiston perusteella lasketut kokonaismatkustajamääräestimaatit ovat pääsääntöisesti laskentalaitteisiin perustuvia estimaatteja pienempiä. Ilmiötä selittää vakiokertoimet, joilla käsinlaskentojen matkustajamäärät yleistetään koko populaatiota koskeviksi estimaateiksi. Viimeisen parinkymmenen vuoden aikana ihmisten liikkumistottumuksissa on tapahtunut suuria muutoksia, joita vanhat 1990-luvulla kehitetyt vakiokertoimet eivät osaa huomioida. Eräs yksittäinen liikkumistottumuksiin positiivisesti vaikuttanut tekijä on vähittäiskauppojen aukioloaikojen laajentaminen etenkin sunnuntaina. Vakiokertoimet asettavat myös haasteita laskentapäivien valintaan. Laskentapäivien tulee edustaa kuukauden keskimääräistä tasoa, mikä on haasteellista saavuttaa matkustajakäyttäytymisen kannalta hyvin vaihtelevissa kuukausissa kuten esimerkiksi joulukuussa.

Matkustajalaskentaprosessi voidaan jakaa karkeasti kahteen pääluokkaan: suunnitteluvaiheeseen ja varsinaiseen tutkimukseen. Suunnitteluvaihe on kertaluonteisesti toteutettava



Kuva 1: *Kaikkien raitiolinjojen matkustajat yhteensä vuonna 2011.*

prosessin määrittely, jossa laaditaan matkustajamäärätutkimuksen eri osavaiheiden toimintaperiaatteet. Suunnitteluvaiheeseen voidaan palata, mikäli tutkimusta halutaan kehittää tai ainakin selvittää sen kehitysmahdollisuudet. Määrittelyn pohjalta toteutetaan toistuvin väliajoin matkustajamäärätutkimus, joka pitää sisällään otoksen poiminnan, tilastollisen estimoinnin sekä tulosten raportoinnin. Automaattisten matkustajalaskentalaitteiden myötä prosessia täytyy kehittää kaikkien edellisten osavaiheiden osalta.

Helsingin seudun bussi-, raitiovaunu-, metro- sekä lauttaliikenteen matkustajamäärätutkimus tehdään kuukausittain, johon junien manuaalilaskenta ei suoritetulla laajuudellaan taipunut. Lähijunaliikenteen automaattisen matkustajalaskennan perimmäinen tarkoitus on nykyaikaistaa matkustajamäärätutkimus luotettavampien sekä ajankohtaisempien matkustajamääräestimaattien saavuttamiseksi. Nykyaikaistaminen tarkoittaa tässä sellaisten matkustajamäärätutkimuksen mahdollisimman automatisoitujen osavaiheiden yhdistelmää, joilla pystytään vastaamaan nykyajan kiivasrytmiseen liikkumistottumusten vaihteluun. Projektissa pyritään automatisoimaan matkustajalaskennan prosessiin liittyvät tehtävät, jotka ovat ennen vaatineet paljon käsityötä ja suuria kustannuksia. Tällaisia ovat esimerkiksi tiedonkeruu sekä aineiston tallentaminen sähköiseen muotoon, jotka on tähän asti tehty täysin käsityönä. Lisäksi raportointia ja tiedonkeruuta halutaan tehostaa luomalla erityinen raportointityökalu, jonka tietoihin sekä HSL:llä että VR:llä on käyttöoikeudet. Aiemmin raportointi suoritettiin julkaisemalla paperisia sekä Excel-pohjaisia raportteja, jotka tuotettiin käyttämällä matkustajalaskentasovelluksen (MATLAS) rapor-

tointimalleja.

Manuaalilaskentojen pohjalta tehtiin useita eri käyttötarkoituksiin suunniteltuja raportteja. Niiden käyttö on hyvin vakiintunutta, joten myös APC-laitteisiin perustuvan tilastollisen estimoinnin tulee mahdollistaa vähintään vastaavanlainen raportointi. Lisäksi raporttien tulee olla yhdenmukaisia keskenään siinä mielessä, että populaation erillisten osajoukkojen kokonaismatkustajamääräestimaattien tulee summautua aina koko populaatiota koskevaksi kokonaisestimaatiksi. Tämä tarkoittaa esimerkiksi sitä, että linjoittaisten kokonaismatkustajamääräestimaattien tulee summautua koko lähijunaliikennettä koskevaksi kokonaisestimaatiksi annetulla aikajaksolla. Yhteneviä estimaatteja perustellaan sillä, että tulosten käyttäjiä on paljon ja he ovat taustaltaan erilaisia. On kiusallista, jos samasta parametrista on liikkellä erilaisia piste-estimaatteja. Lisäksi henkilöt, jotka eivät tiedä estimoinnin takana olevasta tilastollisesta menetelmästä, voivat hämmentyä tai pitää estimaatteja jopa epäluotettavina. Väärinkäsitysten ehkäisemiseksi raporteissa tulee olla mahdollisuus erottaa todellinen tieto estimoidusta tiedosta.

Junayksiköihin asennetut automaattiset matkustajalaskentalaitteet mittaavat nousevat ja poistuvat matkustajat kaikilla pysähdysasemilla. Yksittäisen matkustajan matkan pituus ei ole kuitenkaan aineistosta havaittavissa. Yksittäisten matkojen muodostama jakauma on mallin piilomuuttuja, joka määrittää asemittain havaittavissa olevien kokonaisnousijoiden ja -poistujien jakauman. Tutkielmassa näytetään, että riippumattomien ja samoinjakautuneiden lähtöjen ollessa kyseessä, voidaan löytää yksittäisen matkustajan matkan pituuden todennäköisyysjakauma. Mikäli todennäköisyys poistua kullakin asemalla on sama kaikille matkustajille riippumatta heidän lähtöasemasta, mallilla on Markov-ominaisuus.

2 APC-laitteet lähijunaliikenteessä

2.1 Laskentalaitteet ja aineiston tietokanta

Lähijuniin sijoitetut laskentalaitteet ovat saksalaisen Dilax Intelcom GmbH:n toimittamia. Niiden laskentamenetelmä perustuu aktiivisten infrapunasensoreiden lähettämien säteiden heijastumiskertojen lukumäärään matkustajien osuessa niihin. Infrapunasäteitä on asetettu kaksi peräkkäin, jotta matkustajien kulkusuunta voidaan määrittää osumisjärjestyksen perusteella. Sensorit sijaitsevat junayksiköissä, joissa ne on asetettu ovien yläpuolelle.

Lähijunaliikenteessä käytettävä kalusto käsittää noin 150 junayksikköä, joista keväällä 2013 noin 25 % oli varustettu laskentalaitteilla. Erilaisten kalustosarjojen laskentalaitteelliset osuudet on esitetty taulukossa (1). Sm1- ja Sm2-kalustosarjat ovat vanhimpia ja ne

tullaan osittain korvaamaan uusilla laskentalaitteellisilla Sm5-yksiköillä. Yhteensä kuusi Sm4-yksikköä tullaan varustamaan laskentalaitteilla vuoden 2013 loppuun mennessä, mihin esitetään tarkemmat perustelut kappaleessa (2.2.1). Laskentalaitteellisen kaluston prosentuaalisesta määrästä ei voida suoraan päätellä, että myös otantasuhde olisi esimerkiksi päivän, viikon tai kuukauden aikana samansuuruinen. Usein otantasuhde on VR:n lähiliikenteessä suurempi, koska liikenteessä suositaan uusien Sm5-laskentayksiköiden käyttöä aikoina, jolloin alemman kysynnän vuoksi koko kalustoa ei ole tarvetta pitää liikenteessä.

Laskentalaitteiden keräämät tiedot siirretään FTP:n avulla HSL:n palvelimelle, josta Dilax poimii sen omalle palvelimelleen. Dilax käsittelee raakadataa kohdistamalla ovikohtaiset mittaukset sopiviin linjoihin ja aikatauluihin käyttämällä hyväksi laitteiden keräämiä kellonaikoja, GPS-koordinaatteja sekä etäisyyttä edellisestä pysähdyksestä. [6, 2012.] Raakadatan kohdistaminen oikeille linjoille ja lähdöille on huomattavasti helpompaa junaliikenteessä kuin esimerkiksi bussiliikenteessä, jossa erilaiset reittimahdollisuudet ja liikenteessä olevan kaluston määrä muodostavat vaikeasti hallittavan kokonaisuuden. Lähijunaliikenteessä keskimäärin $94 - 100$ % kaikista raakadatan asemamittauksista saadaan kohdistettua. Joskus luku saattaa olla pienempi, mihin on usein syynä mittausten tiheä vuoroväli. [26, 2012.]

Jalostettu data on saatavissa tietokanta- ja raportointisovelluksessa DavisWeb Mobilessa. Sovelluksen *Filter*-alueella aineistoa voi suodattaa esimerkiksi aikamuuttujien, linjan, suunnan, asemien ja mittaavien yksiköiden mukaan. Alueella voi myös tehdä yksinkertaista datan puhdistamista tiettyjen muuttujien perusteella. Tästä kerrotaan tarkemmin kappaleessa (4.2.1). Suodattamastaan aineistosta käyttäjä voi *Table report*-alueella laskea esimerkiksi otantasuhteita sekä matkustajamäärien keskiarvoja, summia, minimejä ja maksimeja erilaisten taustamuuttujien mukaan. Kappaleessa (4.1.1) perehdytään DavisWebin käyttämään painotusmenetelmään, jolla se laskee myös matkustajamäärien piste-estimaatteja.

Kalustosarja	Laskentayksiköt (lkm)	Kaikki yksiköt (lkm)	Laskentayksiköiden prosentuaalinen osuus
Sm1	0	40	0 %
Sm2	11	50	22 %
Sm4	1	30	3 %
Sm5	24	24	100 %

Taulukko 1: *Laskentalaitteellisen kaluston osuus eri kalustosarjoissa 5/2013.*

2.2 Tutkimusongelmia

2.2.1 Ei-todennäköisyysotantaan perustuva otos

Matkustajamäärätutkimuksen tavoitteena on luoda totaaliestimaatteja koko lähijunaliikenteen populaatioon, joka käsittää kaikki lähijunaliikenteen lähdöt kuukauden aikana. Perusjoukon eri käsitteet lähijunaliikenteessä kuvataan tarkemmin kappaleessa (2.2.3). Jotta estimaatit eivät olisi harhaisia, täytyy tutkittavasta perusjoukosta poimia mahdollisimman edustava otos, eräänlainen pienoismalli. Lähijunaliikenteen populaatio on jakaantunut luonnollisesti homogeenisiin ryhmiin, jotka ovat uniikkeja lähtöjä. Uniikilla lähdöllä on tietyt esimerkiksi linjaan, kellonaikaan, viikonpäivään ja suuntaan sidotut ominaisuudet ja se voidaan operoida useampana päivänä viikon tai kuukauden aikana. Siten linjan A kello 8:09 lähtevä lähtö Helsingistä arkipäivinä on esimerkki uniikista lähdöstä, jonka realisoinnit liikenteessä tuottavat melko samanlaisia tuloksia. Tutkimuksen kannalta olisi ihanteellista saada vähintään yksi näyte kustakin uniikista lähdöstä kuukaudessa. Tämän toteuttaminen harkinnanvaraisella otannalla, puhumattakaan todennäköisyysotannasta, oli epärealistinen tavoite johtuen liikenteen suunnitteluun kohdistuvista haasteista ja suunnitelmien epävarmasta toteutumisesta.

Tutkimuksessa jouduttiin siis tukeutumaan nykyiseen kalustonkierrätysmetodiikkaan, jossa (laskenta)yksiköiden liikkumiseen vaikuttavat mm. kausittainen kalustonkierrätyssuunnitelma sekä erilaiset satunnaistekijät. Kuvitellaan hyvin yksinkertainen esimerkki, jossa kalustolle on luotu tarkka ns. kalustoketju, jonka se kulkee päivän aikana. Satunnaisuutta kalustoketjussa on ainoastaan junayksiköiden sijoittuminen ketjun aloituslähdöille. Tällöin mitattavat lähdöt ovat riippuvaisia siitä, mille aloituslähdölle laskentayksiköt sijoittuivat. APC-laitteiden keräämät näytteet perustuvat siis kalustoketjun määräämään verkosto-otantaan. Todellisuudessa tilannetta mutkistaa muut satunnaistekijät, kuten esimerkiksi sääoloista, laiterikoista ja onnettomuuksista johtuvat poikkeukset liikennesuunnitelmista.

Ei-todennäköisyysotannassa tutkijan on tärkeää tiedostaa, keillä kiinnostuksen alla olevista tutkimusyksiköistä on mahdollisuus tulla poimituksi otokseen. Jotta lähijunaliikenteen lähtö voisi tulla mitatuksi APC-laitteilla, tarvitaan oletus, että laskentalaitteellinen junayksikkö täyttää lähdöllä tarvittavan kaluston ominaisuudet. Esimerkiksi Sm2-yksiköiden nopeus ei riitä Riihimäelle ja Lahteen ulottuvien H-, Z- ja R-linjojen liikennöimiseen. Myös Sm5-yksiköt ovat kyseisillä linjoilla vaikeasti käytettäviä kalustonkiertoon liittyvistä ongelmista johtuen. Näistä syistä edellä mainitut linjat tulevat nykyisellä taulukon (1) mukaisella laskentakalustolla aliedustetuiksi. Niiden edustavuutta halutaan kasvattaa varusta-

malla osa Sm4-yksiköistä APC-laitteilla. Joitakin vuoroja ajetaan veturivetoisilla junilla, joissa ei ole laskentalaitteita. Ne kuuluvat siis alipeittoon (*under-coverage*), jonka suuruus on kuitenkin tiedossa. Alipeittoon reagoidaan siirtämällä veturivetoiset lähdöt automaattisen matkustajamäärätutkimuksen ulkopuolelle ja arvioimalla niiden matkustajamääriä manuaalilaskennoin.

Matkustajamäärätutkimuksen tutkimusyksikkö on lähtö. Vastauskato (*non-response*) voidaan jakaa sekä yksikkö- että erävastauskatoon (*unit and item non-response*). Yksikkövastauskato käsittää lähdöt, joista ei ole saatu mitään tietoa. Liikennöiviä junia muodostetaan yhdestä tai useammasta junayksiköstä, joten on mahdollista, että osa junasta pitää sisällään laskentalaitteet ja osa ei. Tälläinen osittainen tiedonsaanti tutkimusyksiköstä on erävastauskatoa. Kalustosuunnitelmista johtuen osalla tutkimusyksiköistä on muita suurempi todennäköisyys tulla sisällytetyksi APC-laitteiden keräämään näytteeseen. Suuremman vastaustaipumuksen omaavat esimerkiksi lähdöt, joilla suunniteltu kalustosarja on Sm5. Aineistosta laskettaviin estimaatteihin syntyy harhaa, jos tulosmuuttuja korreloi vastaustaipumuksen kanssa [11, s.2, 2009]. Tutkielman keskeiseksi ongelmaksi muodostui vastauskadosta johtuvan vinouman oikaiseminen, johon kappale (3) tarjoaa kaksi eri tilastollista menetelmää.

2.2.2 Erävastauskato

Edellisessä kappaleessa kuvattiin, miten erävastauskatoa esiintyy lähijunaliikenteessä. On selvää, että matkustajat eivät jakaudu tasaisesti junayksiköiden välillä. Syitä tähän ei aleta tässä sen kummemmin pohtimaan. Jotta kuormitusten erot voitaisiin ottaa huomioon mahdollisesti eri linjoilla, päivinä ja vuorokauden aikoina, tarvitaan tieto lähdön yksikkökokoonpanosta eli operoivien junayksiköiden yksilöintitunnuksista sekä niiden järjestyksestä toisiinsa nähden. Tietoa ei ole tällä hetkellä automaattisesti saatavissa, joten kuorman variaation huomioon ottaminen oli tässä vaiheessa liki mahdotonta.

Erävastauskatoa muodostavat myös APC-laiteviat, joten ainoastaan kokonaisten laskentajunien muodostaminen ei tulisi ratkaisemaan erävastauskadon ongelmaa. Tilannetta helpottaa kuitenkin huomattavasti se, että laskentayksiköiden sijoittuminen junassa sekä APC-laitevikojen ilmaantuminen ovat kutakuinkin satunnaisia tapahtumia. Satunnaisuus takaa sen, että pitkällä aikavälillä näytteitä saadaan tasaisesti eripuolilta junaa.

Tulevaisuudessa uusi IVU-järjestelmä tuo mukanaan uutta automaattisesti käytettävissä olevaa dataa sisältäen mm. toteutuneen aikataulun ja yksikkökokoonpanon. Matkustajamäärätutkimuksen kehityspolkua on IVU:n käyttöönoton yhteydessä arvioitava uudelleen.

2.2.3 Kehikkovirheet

Toteutunut liikenne ei aina vastaa suunniteltua; erityisesti haastavat talviolosuhteet myöhästyttävät vuoroja ja aiheuttavat kalustomuutoksia. Tutkielman kiinnostusperusjoukko (*population of interest*) sisältää kaikki ajetut lähdöt sekä niiden toteutuneen aikataulun ja kaluston. Käytettävissä oleva kehikkoperusjoukko (*frame population*) perustuu kausittaisiin aikataulu- ja kalustosuunnitelmiin. Siitä ei ole saatavilla automaattisesti tuotettua päivitystä eli ns. päivitettyä kehikkoperusjoukkoa (*updated frame population*), joten lähtötietojen muuttaminen vaatisi käsityötä.

Peruutettujen lähtöjen muodostama ylipeitto (*over-coverage*) on eräs kehikkovirheen (*frame error*) lähde, joka syntyy, kun kehikkoperusjoukkoa ei päivitetä. Vuonna 2011 peruutettuja lähtöjä oli kuukaudessa keskimäärin 0,92 % [2, 16, s.43, 2012]. Keskimääräinen suunniteltujen lähijunaliikenteen lähtöjen lukumäärä kuukaudessa oli kyseisenä vuonna 24 000, mikä tarkoittaa, että keskimäärin 220 lähtöä peruttiin kuukausittain. Täsmällisyys on kuitenkin lähtenyt haastavien talvien 2009-2010 ja 2010-2011 jälkeen nousuun. Peruutettujen lähtöjen saaminen muotoon, joka on yhdistettävissä mikrodatan kanssa, vaatii paljon käsityötä. Siitä syystä ylipeitosta kärsitään siihen asti, kunnes tieto saadaan automaattisesti.

Kolmas kehikkovirheen lähde on junayksiköiden lukumäärä kullakin lähdöllä. Virhe syntyy, kun osittain tai kokonaan lasketun lähdön toteutuneiden yksiköiden lukumäärä poikkeaa suunnitellusta. Jos yksiköitä on vähemmän kuin on suunniteltu, virhe vaikuttaa matkustajamääriin positiivisesti. Vastaavasti virhe on negatiivisvaikutteinen yksikkömäärän ollessa suurempi kuin mitä on suunniteltu. Vuonna 2011 vajavaisten lähtöjen lukumäärä HSL-liikenteessä oli 997, mikä vastaa noin 0,4 prosenttia kaikista lähdöistä. Vajavaisia lähtöjä esiintyi suhteellisesti eniten Sm1/Sm2-kalusteisilla lähdöillä, joten vajavaisten lähtöjen prosenttiosuus APC-laitteisilla lähdöillä on vieläkin pienempi. Ylimääräisellä rungolla varustettuja lähtöja esiintyy erittäin harvoin, ja esiintyessäänkin ylimääräinen yksikkö usein suljetaan pois käytöstä [10, 2012].

Tavoiteperusjoukko (*target population*) on sellainen realistisesti tavoitettavissa oleva joukko, jota todella yritetään tutkia. Tämän kappaleen sekä kappaleen (2.2.1) yhteenvetona voidaan esittää, että matkustajamäärätutkimuksen tavoiteperusjoukko on tutkittavan kuukauden kaikki lähijunaliikenteen suunnitellut lähdöt, joista on poistettu veturivetoiset lähdöt. Ulkopuolelle rajautuu aluksi myös linjat H, Z ja R, kunnes tarvittava määrä Sm4-yksiköitä saadaan varustettua laskentalaitteilla. Matkustajamäärien tuloksia tullaan siis esittämään tämän edellä kuvatun tavoiteperusjoukon tasolle.

2.3 APC-mittausten vertaaminen käsinlaskentoihin

Sm5-junille suoritettiin manuaalinen tarkistuslaskenta helmi- ja maaliskuussa 2013. Käsinlaskentoja verrattiin matkustajalaskentajärjestelmän tuottamiin arvoihin mittaustarkkuuden selvittämiseksi. Manuaalilaskentoja suoritti vähintään 12 henkilön ryhmä, joka liikkui kentällä ennakkoon tehdyn työohjelman mukaisesti. Jokaiselle ovelle sijoittui aina kaksi laskijaa, joista toinen laski sisään tulevat ja toinen ulos menevät matkustajat. Tulokset kirjattiin paperiselle lomakkeelle, josta ne myöhemmin sähköistettiin Excel-taulukkoon.

Kaikkiaan laskentatapahtumia eli asemamittausten lukumäärä oli nousijoille 694 ja poistujille 693. Käsinlaskentanäytteitä jouduttiin hylkäämään esimerkiksi puuttuvan tai epäselvän merkinnän, häiriötekijöiden sekä vastaavan APC-mittauksen puuttuneisuuden takia. Lisäksi ns. varmat tapaukset, kuten nousijat pääteasemalla, jätettiin tuloslaskelmien ulkopuolelle. Hyväksyttyjä näytteitä saatiin kerättyä 18 eri yksiköstä, kun tutkimusaikana liikenteessä oli kaikkiaan 23 yksikköä.

Tarkastellaan asemalla tapahtuvien nousija- ja poistujamittausten poikkeamia manuaalilaskennasta. Manuaalisesti sekä laskentalaitteilla lasketuista lähdöistä tuotetaan myös asemien väliset kuormat, joiden poikkeamat otetaan myös tarkastelun alle. Määritellään siis muuttujat DON, DOFF ja DLOAD kuten

$$\begin{aligned} \text{DON} &= \text{nousijat(APC)} - \text{nousijat(MANUAALI)} \\ \text{DOFF} &= \text{poistujat(APC)} - \text{poistujat(MANUAALI)} \\ \text{DLOAD} &= \text{kuorma(APC)} - \text{kuorma(MANUAALI)}. \end{aligned}$$

Muuttujan DON keskiarvo oli 0,11, mikä tarkoittaa, että laskentalaitteet laskivat asemalla keskimäärin 0,11 matkustajaa enemmän. Keskiarvo muuttujalle DOFF oli 0,26 ja muuttujalle DLOAD -0,96. Kuorma on siis keskimäärin ollut pienempi verrattuna manuaalilaskentaan, jolloin matkustajakilometrejä jää uupumaan. Tämä johtuu siitä, että muuttujan DOFF arvot ovat niin suuria, että ne kumoavat jopa keskimäärin positiiviset muuttujan

							95 % luottamusväli	
	n	ka	s.d.	min	max	p-arvo	alaraja	yläraja
DON	694	0,11	1,78	-9	6	0,1035	-0,02	0,24
DOFF	693	0,26	1,81	-8	12	0,0002	0,12	0,39
DLOAD	709	-0,96	6,15	-23	18	0,0000	-1,41	-0,50

Taulukko 2: *Muuttujien DON, DOFF ja DLOAD tunnuslukuja.*

DON arvot. Se, että laitteet laskevat poistujia enemmän kuin nousijoita, on muiden vastaavien tutkimusten kanssa samansuuntainen [12, s.75, 2010].

Todennäköisyyslaskennan keskeisen raja-arvolauseen perusteella otoskeskiarvo noudattaa likimain normaalijakaumaa riippumatta siitä, miten tutkittava muuttuja on jakautunut. Poikkeamien keskiarvoja tarkasteltaessa huomataan, että eroja manuaalilaskennan ja APC-laitteiden välillä on syntynyt. Mielenkiintoista on se, että ovatko erot tilastollisesti merkitseviä. Muuttujille DON, DOFF ja DLOAD tehtiin nollahypoteesin testaus, jossa otoskeskiarvoille laskettiin 95 % luottamusväli. Mikäli nolla kuuluu luottamusvälille, ei aineisto anna tarpeeksi evidenssiä nollahypoteesin hylkäämiseen. Jos taas nolla ei kuulu luottamusvälille, nollahypoteesi hylätään ja voidaan sanoa, että poikkeamat tutkittavassa muuttujassa ovat systemaattisia. Tulokset näyttävät, että muuttujien DOFF ja DLOAD osalta nollahypoteesi tulee hylätyksi (2). Laskentalaitteet siis laskevat poistujia systemaattisesti liikaa, mistä syystä kuorma on systemaattisesti pienempi.

Eräs tuloksiin vaikuttava tekijä on konduktööri, joka usein kurkkii ovesta ulos ja antaa kuljettajalle merkin kun ovet voi sulkea ja matka jatkua. Tilanteessa konduktööri ei selkeästi ylitä laskenta-aluetta, vaan oleskelee oviaukossa suoraan laskentasensoreiden alapuolella. Laskijoita ohjeistettiin olemaan laskematta oviaukossa oleskelevia henkilöitä, jotka Dilaxin mukaan ovat liian epävarmoja tilanteita laskentalaitteiden päätettäväksi [7, 2013]. Kuitenkin tarkasteltaessa aineistoa erikseen niiden asemien osalta, jossa konduktööri on käynyt ovella ja niiden joissa konduktööri ei ollut käynyt ovella, havaitaan poikkeamien keskiarvoissa selkeä ero. Konduktöörin kurkkaukset sisältävien asemamittausten keskiarvo oli muuttujalle DON 0,393 ja muuttujalle DOFF 0,453. Konduktöörin kannalta häiriöttömille asemittauksille vastaavat arvot olivat -0,006 ja 0,175.

3 Vastauskadon oikaisumenetelmiä

3.1 Painotusmenetelmät

Painotusta käytetään usein osana estimointimenetelmää, kun poimittu otos ei kata koko tavoiteperusjoukkoa, mutta tuloksia halutaan esittää koko tavoiteperusjoukon tasolle [14, s.108, 2010]. Painotuksella pyritään kompensoimaan vastauskadosta ja otanta-asetelmasta johtuvaa otosaineiston vinoumaa, jotta tämä saadaan paremmin vastaamaan tavoiteperusjoukkoa. Koska vastauskatoa esiintyy empiiris-kvantitatiivisissa tutkimuksissa lähes aina, painotusta tarvitaan usein myös silloin, kun koko tavoiteperusjoukko on poimittu otokseen.

Lisätieto (*auxiliary infromation*) on otannan ulkopuolista tietoa tutkimusyksiköistä tai

näiden kokonaismäärästä ja se liitetään otantatiedostoon identifikaatiomuuttujien avulla. Lisätietoa voidaan sisällyttää otanta- tai estimointiasetelmaan tai vaihtoehtoisesti molempiin riippuen siitä, millä strategialla uskotaan saavan tehokkainta estimointia. Sopivalla lisätiedon käytöllä estimaattien tehokkuutta voidaan kasvattaa merkittävästi [15, s.4, 2004]. VR:n lähiliikenteessä apumuuttujat ovat muuttujat linja, suunta, lähtöaika, päivämäärä, viikonpäivä sekä suunniteltu yksikkömäärä. Kaikki muuttujat ovat käytettävissä estimointiasetelmassa.

3.1.1 Peruspainot

Todennäköisyysotantaan perustuvissa tutkimuksissa ensimmäiset painot eli ns. asetelmapainot (*design weight*) lasketaan brutto-otokselle. Asetelmapainojen muodostaminen perustuu tutkimusyksiköiden todennäköisyyteen sisältyä otokseen. Olkoon π_k tutkimusyksikön k otanta-asetelman mukainen todennäköisyys sisältyä otokseen. Asetelmapaino saadaan sisältymistodennäköisyyden käänteislukuna:

$$a_k = \frac{1}{\pi_k}.$$

Sanallisesti se ilmaisee, kuinka montaa yksikköä otostutkimusyksikkö k edustaa tavoiteperusjoukossa. On syytä huomata, että asetelmapainot summautuvat tutkimusyksiköiden lukumääräksi [14, s.57, 2010].

Vastauskadon esiintyessä asetelmapainot on muunnettava koskemaan netto-otosta. Uusia painoja kutsutaan peruspainoiksi (*base weight*), ja ne saadaan korvaamalla otoskoko n vastaajien määrällä r. Mikäli lisätietoa on vielä käytettävissä, painoja voidaan parantaa edelleen esimerkiksi kalibroimalla, vastaustodennäköisyyksiin perustuvalla menetelmällä tai jälkiosittamalla. Jälkimmäinen menetelmä esitetään tarkemmin seuraavassa kappaleessa. Uudelleenpainotusta voidaan käyttää myös ei-todennäköisyysotannan aineistoihin, joissa alkioiden sisältymistodennäköisyyttä ei tiedetä.

Esimerkki 3.1. Oletetaan, että perusjoukko koostuu N:stä alkiosta, joista otokseen poimitaan n alkiota yksinkertaisella satunnaisotanalla.

(i) Tutkimusyksikön k asetelmapaino on

$$a_k = \frac{1}{\pi_k} = \frac{N}{n}.$$

(ii) Vastaava peruspaino saadaan vaihtamalla otoskoko n vastaajien määräksi r, jolloin

$$w_k = \frac{N}{r}.$$

3.1.2 Jälkiosittaminen

Jälkiosittaminen (*post-stratification*) tarkoittaa menetelmää, jossa otanta-asetelmassa laadittujen ositteiden sisälle muodostetaan uusia ositteita. Lisäinformaation tulee olla luokiteltua, jotta sitä voidaan käyttää jälkiositteiden muodostamisessa, ja sen tulee olla tiedossa kaikilta kehikkoperusjoukon alkioilta. Eräs jälkiosittamisen päämäärä on vastaustodennäköisyyksiltään mahdollisimman homogeenisten ryhmien muodostaminen, jonka jälkeen puuttuneisuusmekanismin voidaan olettaa olevan kunkin ositteen sisällä satunnainen eli MARS (*missing at random under sampling design*). Uudet painot muodostetaan alkuperäisten ositteiden sisälle käyttäen hyväksi päivitettyä kehikkoperusjoukkoa, mikäli se on saatavilla. Jälkiositteita muodostettaessa on huomioitava, että kaikista ositteista on olemassa tarpeeksi havaintoja. Mikäli ositteesta ei ole olemassa yhtään havaintoa, on uudet painot kaikille ositteen alkioille nolla, milloin niitä ei voida sisällyttää tavoiteperusjoukkoon.

Esimerkki 3.2. Olkoon aineisto poimittu ositetulla otannalla, jossa ositteen i, $i = 1, ..., I$, sisältä on poimittu n_i alkiota yksinkertaisella satunnaisotannalla. Aineistoa on päätetty jälkiosittaa, jolloin ositteen i sisälle muodostetaan uudet ositteet ij, $j = 1, ..., J$. Jälkiositteiden perhettä merkitään kirjaimella U ja se voidaan ilmaista matemaattisesti, kuten

$$U = \{ij \mid i = 1, ..., I\,;\, j = 1, ..., J\}\,.$$

Uusi paino tutkimusyksikölle k on

$$w_k^* = \sum_{ij \in U} 1_{[k \in ij]} \frac{N_{ij}}{r_{ij}},$$

missä N_{ij} on perusjoukon koko ja r_{ij} vastaajien lukumäärä jälkiositteessa ij. Funktio $1_{[k \in ij]}$ on indikaattorifunktio, joka saa arvon 1, kun $k \in ij$, ja arvon 0 muulloin.

Toisin kuin esimerkissä (3.2) jälkiositteiden lukumäärä esiositteen sisällä voi vaihdella eikä kaikkia esiositteita edes tarvitse jälkiosittaa. Luokkien yhdisteleminen on suositeltavaa ainakin silloin, kun havaintoja on niukasti. Jälkiositus voidaan tehdä myös tilanteessa, jossa mitään esiositteita ei ole tehty. Tällöin ositteiden muodostaminen on täysin vapaata olemassa olevan lisäinformaation puitteissa. Jälkiosittamisen eräs etu on se, että erilaisten vastaajaryhmien reunajakaumat saadaan painotettua vastaamaan populaation rakennetta. Siten jokaiselle jälkiositteelle saadaan oma estimaatti. Tämä on suotuisa ominaisuus VR:n lähiliikenteessä, jossa estimaatteja tarvitaan erilasten taustamuuttujien mukaan.

3.2 Imputointi

Imputoinnilla korvataan puuttuvaa, virheellistä tai epäyhdenmukaista tietoa keinotekoisesti tuotetuilla korvikearvoilla. Imputointia kannattaa käyttää, jos estimaatin laadun uskotaan paranevan verrattuna siihen mitä ilman imputointia olisi saavutettu [14, s.121, 2010]. Usein korvikearvo perustuu ei-puuttuvien arvojen avulla tehtävään tilastolliseen estimointiin. Useita imputointimenetelmiä on esitetty ja niiden käyttö riippuu oleellisesti aineiston tyypin, laadun ja puuttuvien havaintojen mukaan [23, s.74, 2007]. Imputointia käytetään usein erävastauskadon täydentämiseen, mutta myös yksikkövastauskatoa voidaan imputoida. Tällöin puhutaan massaimputoinnista (*mass imputation*).

Joissakin tapauksissa puuttuva tieto voidaan päätellä varmasti vastaajan muiden arvojen perusteella. Mikäli tällainen looginen imputointi (*deductive imputation*) on mahdollista, tulisi sitä soveltaa aina ensisijaisesti. Joskus menetelmää käytetään myös silloin, kun arvot voidaan päätellä hyvin suurella todennäköisyydellä [24, s.225, 2011]. Muut imputointimenetelmät voidaan jakaa karkeasti luovuttajan tyypin mukaan malli- ja vastaajaluovuttajaperusteisiin. Malliluovuttajaperusteinen imputointi (*model-donor imputation*) perustuu vastaajien aineistosta estimoituun malliin, joka tuottaa imputoidut arvot joko deterministisesti tai stokastisesti. Deterministinen imputointi tuottaa aina saman arvon, kun taas stokastinen imputointi sisältää satunnaisuutta. Vastaajaluovuttajaperusteisessa imputoinnissa (*real-donor imputation*) arvot lainataan samankaltaisilta vastaajilta. Tämän menetelmän etu on se, että arvot ovat aina todellisuudessa mahdollisia.

Aineiston ominaisuuksien lisäksi imputointimenetelmää valittaessa on huomioitava imputoinnille asetettu tavoitetaso. Tavoite voi olla esimerkiksi imputoitavien arvojen saaminen mahdollisimman lähelle oikeita arvoja tai aineiston jakauman saaminen lähelle todellista jakaumaa. Mikäli käyttäjän intressi on yksinkertaisissa estimaateissa, kuten kokonais- ja keskiarvoestimaateissa, ei jakauman uskottavuuteen tarvitse kiinnittää erityistä huomiota. Kuitenkin jakaumatason estimoinnin onnistuminen tarkoittaa aina myöskin aggregaattitason estimoinnin onnistumista [14, s.125, 2010].

Malliluovuttajaperusteisista malleista regressiomallit ovat kaikkein yleisimmin käytettyjä [24, s.225, 2011]. Apumuuttujien x_i, $i = 1, 2, \ldots, p$, avulla muodostetaan vastaajien aineistosta malli, joka ennustaa tulosmuuttujan y arvoja. Esimerkiksi lineaarinen regressiomalli on muotoa

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_p x_p + \varepsilon, \tag{3.1}$$

missä $\alpha, \beta_1, \beta_2, \ldots \beta_p$ ovat mallin parametrit ja ε on residuaali eli virhetermi. Kun mallin

parametrit muutetaan estimaateiksi, saadaan ennustearvo

$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots \widehat{\beta}_p x_p.$$

Mikäli jakauman halutaan olevan lähellä totuutta, satunnaisluku ε on hyvä sisällyttää malliin. Sen pois jättäminen pienentää keskihajontaa; jakaumasta tulee monihuippuinen ja ohuthäntäinen. Aggregaattitason tavoitteen ollessa kyseessä satunnaistermi on kuitenkin turha. Se voi jopa vääristää tuloksia, jos sen odotusarvo ei ole nolla. [24, s.231, 2011.] Imputoitu arvo tutkimusyksikön i muuttujalle y voidaan siis muodostaa kahdella vaihtoehtoisella tavalla:

$$\begin{aligned} \widetilde{y}_i &= \widehat{y}_i = \widehat{\alpha} + \widehat{\beta}_1 x_{i_1} + \widehat{\beta}_2 x_{i_2} + \ldots \widehat{\beta}_p x_{i_p} \\ \widetilde{y}_i &= \widehat{y}_i + \varepsilon_i = \widehat{\alpha} + \widehat{\beta}_1 x_{i_1} + \widehat{\beta}_2 x_{i_2} + \ldots \widehat{\beta}_p x_{i_p} + \varepsilon_i. \end{aligned} \tag{3.2}$$

Satunnaistermi ε valitaan usein joko tasaisesta $(0, 1)$-jakaumasta tai normaalijakaumasta odotusarvolla nolla ja aineistosta lasketulla keskihajonnalla [14, s.133, 2010]. Seuraavaksi esitettävät keskiarvo-, mediaani ja suhdeimputointimenetelmät käsitellään vain tapauksessa, jossa satunnaistermiä ei sisällytetä malliin. Tarvittaessa satunnaistermin lisääminen tapahtuu samoin kuin regressioimputoinnissa.

Regressiomallien yksi erikoistapaus on keskiarvoimputointi (*mean imputation*). Kun yhtälöön (3.1) ei oteta mukaan apumuuttujia, se supistuu muotoon $y = \mu + \varepsilon$, missä μ on muuttujan y odotusarvo. Tällöin yhtälö (3.2) muuntuu muotoon $\widetilde{y}_i = \widehat{\mu} = \overline{y}_{\text{obs}}$, missä $\overline{y}_{\text{obs}}$ on keskiarvo havaituista muuttujan y arvoista:

$$\overline{y}_{\text{obs}} = \frac{\sum_{k \in \text{obs}} y_k}{r}.$$

Jos otanta-asetelma on monimutkainen, havaintoja voidaan painottaa edellisestä kappaleesta tutuilla painoilla. Olkoon w_i paino tutkimusyksikölle i. Tällöin keskiarvoimputoinnin kaava on

$$\widetilde{y}_i = \overline{y}_{\text{obs}}^{(w)} \equiv \frac{\sum_{k \in \text{obs}} w_k y_k}{\sum_{k \in \text{obs}} w_k}.$$

Kun keskiarvoimputointi suoritetaan homogeenisen ryhmän sisällä, puhutaan keskiarvoimputoinnista imputointisolun sisällä. Ryhmän olisi hyvä olla homogeeninen imputoitavan muuttujan suhteen sekä vastauskadoltaan mahdollisimman satunnainen [14, s.126, 2010]. Keskiarvoimputointi suoritetaan imputointisolun h sisällä seuraavasti:

$$\widetilde{y}_{hi} = \overline{y}_{h;\text{obs}} \equiv \frac{\sum_{k \in h \cap \text{obs}} y_k}{r_h},$$

missä r_h on vastaajien määrä imputointisolussa h.

Keskiarvoimputointia suositellaan käytettävän, jos apumuuttujia on vähän tai niillä ei ole merkittävää vaikutusta muuttujan y arvoon [24, s.247, 2011]. Menetelmä vaatii oletuksen vastaamattomuuden satunnaisuudesta ainakin imputointisolun sisällä. Muissa tapauksissa menetelmä saattaa johtaa harhaisiin estimaatteihin. Jakauman kannalta keskiarvoimputointi on huono, mutta totaalien ja keskiarvojen estimoimiseen se on kuitenkin käyttökelpoinen. Menetelmän soveltaminen imputointisolun sisällä tuottaa vähemmän piikikkään jakauman, koska ryhmien välinen variaatio on otettu huomioon. Ryhmien sisäinen varianssi on kuitenkin liian liberaali. On hyvä huomioida, että keskiarvoimputointi imputointisolujen sisällä tuottaa samat totaalit ja keskiarvot kuin jälkiosituksella saadut estimaatit, jos imputointisoluina käytetään jälkiositteita [24, s.248, 2011].

Keskiarvoimputoinnin lähisukulainen on mediaani-imputointi (*median imputation*), jossa imputoitu arvo $\widetilde{y}_i$ saadaan muuttujan y havaintojen mediaanina joko koko aineiston sisällä, kuten

$$\widetilde{y}_i = \text{median}\left\{y_k \,|\, k \in \text{obs}\right\},$$

tai imputointisolun h sisällä, kuten

$$\widetilde{y}_{hi} = \text{median}\left\{y_k \,|\, k \in h \cap \text{obs}\right\}.$$

Jos imputoitavan muuttujan jakauma on vinoutunut tai imputoitavien arvojen robustisuus halutaan varmistaa, on mediaani-imputointi parempi vaihtoehto kuin keskiarvoimputointi [1, s. 639-647, 2004].

Toinen regressioimputoinnin erikoistapaus on suhdeimputointi (*ratio imputation*), jota voidaan käyttää silloin, kun tulosmuuttujan y suhde apumuuttujaan x on likimain vakio tavoiteperusjoukossa. Menetelmä perustuu yhden selittävän tekijän lineaariseen regressiomalliin, jossa ei ole regressiovakiota. Yhtälö (3.1) muuntuu siis muotoon $y = Rx$, jolloin imputoidut arvot saadaan yhtälöstä $\widetilde{y}_i = \widehat{R}x_i$. Kerroin R on muuttujan y ja apumuuttujan x välinen suhde. Sen estimaatti saadaan havaintojen avulla seuraavasti

$$\widehat{R} = \frac{\sum_{k\in\text{obs}} y_k}{\sum_{k\in\text{obs}} x_k}.$$

Vastaajaluovuttajaperusteisissa imputointimenetelmissä puuttuvan havainnon sisältävälle tutkimusyksikölle i etsitään vastaajien joukosta mahdollisimman samankaltaisia ominaisuuksia omaava luovuttaja (*donor*) d siten, että valittujen ominaisuuksien uskotaan korreloivan muuttujan y kanssa. Valitun luovuttajan arvo y_d imputoidaan tutkimusyksikölle i:

$$\widetilde{y}_i = y_d.$$

Jos yksikkö i sisältää useamman kuin yhden puuttuvan tiedon, usein samaa luovuttajaa käytetään kaikkien tietojen täydentämiseen [24, s.249, 2011]. Luovuttaja voidaan löytää kahdella vaihtoehtoisella tavalla: käyttämällä imputointisoluja tai läheisyyden metriikkaa. Edellisessä kaikki tutkimusyksiköt imputointisolun sisällä oletetaan olevan yhtä mahdollisia, jolloin luovuttaja valitaan niistä satunnaisesti. Tämän ns. *hot-deck imputation*-menetelmän lisäksi on olemassa myös *cold-deck imputation*-menetelmä, jossa imputoitu arvo lainataan toisesta aineistosta, joka voi olla esimerkiksi pitkittäistutkimuksissa edellisen ajanhetken aineisto. Usein imputoitu arvo paranee, mikäli sitä kerrotaan trendikertoimella. Tämä on tunnetusti suhdeimputointia. [24, s.250, 2011.]

Jos on käytettävissä paljon apumuuttujia, voidaan niiden avulla muodostaa läheisyyden mitta joko painottamalla tiettyjä muuttujia tai laittamalla kaikki samanarvoisiksi. Minimoimalla etäisyys löydetään läheisin arvo, jota käytetään imputoitavana arvona. Mikäli useammalla tutkimusyksiköllä on tämä sama arvo, valitaan luovuttaja niistä satunnaisesti. Läheisyysmetriikkana voidaan käyttää myös malliluovuttajaperusteista imputointimallia, jossa arvot lasketaan sekä vastaajille että vastaamattomille [14, s.128, 2010]. Yleisesti läheisyysmittaan perustuvaa imputointia käytetään tilanteissa, joissa apumuuttujat ovat pääosin numeerisia. Muutoin apumuuttujien kategorisoimisella imputoinnin läpiviemiseksi menetettäisiin paljon informaatiota [24, s.250, 2011].

Pitkittäisaineiston imputointi eroaa usein poikkileikkausaineiston imputoinnista. Samasta tutkimusyksiköstä on olemassa aiempaa tietoa, jolloin imputoinnissa ei tavallisesti käytetä muiden yksiköiden arvoja. Usein käytetty menetelmä on ns. *last value carried forward*, jossa imputoitu arvo lainataan yksinkertaisesti edelliseltä ajanjaksolta. Jos estimaattien tuottamisessa ei ole kiire, voidaan imputointi suorittaa myös interpoloimalla eli käyttämällä tutkimusyksikön sekä mennyttä että tulevaa tietoa. Kuukausittaisen lähijunaliikenteen matkustajamääräraportoinnin tulee olla suoritettuna $2-3$ viikon kuluttua kuun loppumisesta, joten seuraavien kuukausien arvoja ei voida hyödyntää. Kuitenkin estimoitaessa jälkikäteen vuoden 2012 kuukausittaisia matkustajamääriä VR:n lähiliikenteessä voitaisiin imputoitavia arvoja etsiä myös seuraavilta kuukausilta.

4 Matkustajamäärätutkimus lähijunaliikenteessä

4.1 Vastauskadon oikaisumenetelmän sekä raportointisovelluksen valinta

Matkustajamäärätutkimuksen suunnitteluvaiheen eräs keskeinen tehtävä oli valita mittaamattomien lähtöjen korvaamisessa käytettävä menetelmä. Usein yksikkövastauskatoa kompensoidaan painottamalla ja erävastauskatoa imputoinnilla, mutta käytännöstä voidaan myös poiketa [21, s.164, 2009]. Aineiston ominaisuuksien lisäksi menetelmän valinnassa tuli ottaa huomioon mahdollisuudet luoda laajennetusta datasta esimerkiksi julkisen liikenteen kehityksen seurantaan sekä junavuorojen ja kalustonkäytön suunnitteluun tarvittavat raportit. Tehokasta raportointia varten tarvitaan erityinen tiedonhallinta- ja raportointijärjestelmä, joka tuottaa raportit automaattisesti käyttäjän asettamien valintojen mukaan. Koska raportoinnin mahdollisuudet sekä raportointisovelluksen valinta ovat riippuvaisia vastauskadon oikaisumenetelmästä, nousi datan käyttöulottuvuus merkittävään rooliin menetelmän valinnassa.

4.1.1 DavisWebin painotusmenetelmät

Dilaxin raportointityökalu DavisWeb sisältää painotusmenetelmiin perustuvan työkalun, jolla matkustajamääristä voidaan tuottaa aggregaattitason estimaatteja. Tutkimusyksikönä toimii lähtö ja ositteina toimivat uniikit lähdöt. Olkoon uniikin lähdön k perusjoukon koko N_k ja vastaavasti näytteiden lukumäärä n_k. Peruspaino on silloin ositteen k alkioille

$$w_k = \frac{N_k}{n_k}.$$

Olkoon edelleen $b_1(k), b_2(k), \ldots b_{n_k}(k)$ ositteen k mitattujen lähtöjen kokonaisnousijamäärät. Tällöin nousijamäärän kokonaisestimaatti on

$$\sum_{k=1}^{U}\sum_{i=1}^{n_k}\frac{N_k}{n_k}\cdot b_i(k) = \sum_{k=1}^{U} N_k \cdot \frac{\sum_{i=1}^{n_k} b_i(k)}{n_k} = \sum_{k=1}^{U} N_k \bar{b}_k \tag{4.1}$$

missä U on uniikkien lähtöjen lukumäärä ja $\bar{b}_k$ mitattujen nousijoiden keskiarvo ositteessa k. Mikäli vähintään yhdestä uniikista lähdöstä ei ole olemassa mittausta, yhtälöstä (4.1) saatu estimaatti ei kata koko tavoiteperusjoukkoa. DavisWebissä tyhjien solujen ongelma on ratkaistu laskemalla korjauspaino

$$w' = \frac{\sum_{k=1}^{U} N_k}{\sum_{k=1}^{U} 1_{[n_k \neq 0]} N_k}, \tag{4.2}$$

joka sanallisesti tarkoittaa suunniteltujen lähtöjen lukumäärän suhdetta ei-tyhjien solujen suunniteltujen lähtöjen lukumäärään. Kertomalla yhtälö (4.1) korjauspainolla saadaan

tavoiteperusjoukon kokonaisestimaatti

$$w' \cdot \sum_{k=1}^{U} N_k \bar{b}_k. \tag{4.3}$$

Kun aineistossa ei ole mittaamattomia uniikkeja lähtöjä, korjauspaino (4.2) on 1, jolloin kokonaisestimaatit (4.1) ja (4.3) ovat samat. Korjauspainon käyttö vaatii oletuksen, että tyhjät solut muistuttavat jakaumaltaan ei-tyhjiä soluja. Tämä tarkoittasi sitä, että nousijamääriltään samantyyppisiä lähtöjä esiintyisi suhteellisesti saman verran sekä tyhjissä että ei-tyhjissä soluissa. Mikäli näin voitaisiin olettaa olevan, ositteet olisi täytynyt muodostaa satunnaisotanalla. Tämä taas on ristiriidassa ositteiden perimmäisen käyttötarkoituksen kanssa.

On mahdollista, että uniikki lähtö liikennöidään sekä arkena että viikonloppuna tai sekä Helsinkiin että Helsingistä päin mentäessä. DavisWebin raportointialueella aineistoa voi jälkiosittaa esimerkiksi suunnan ja viikonpäivän mukaan, jolloin ositteista saadaan entistä homogeenisempia.

4.1.2 Painotusmenetelmät vai imputointi?

Matkustajamäärätutkimuksen suunnitteluvaiheessa oli esillä kolme erilaista vastauskadon oikaisumenetelmän, analyysityökalun sekä raportointityökalun yhdistelmää. Vaihtoehdot on esitetty taulukossa (3). SAS (*Statistical Analysis System*) on tilastollisen aineistonkäsittelyn sekä analysoinnin ohjelmisto, jossa aineistoa käsitellään erityisellä SAS-kielellä. Mikäli sitä päätetään käyttää tilastollisen estimoinnin työkaluna, tarvitaan raportointia varten kehittää uusi SAS-ohjelmiston tulosaineistoa käyttävä ohjelmisto.

DavisWebin käyttö sekä vastauskadon oikaisun että raportoinnin työkaluna oli houkutteleva vaihtoehto, koska toiminnot voitaisiin suorittaa samanaikaisesti ilman aineiston siirtoa sovelluksesta toiseen. Suurimmaksi ongelmakohdaksi nousi kuitenkin DavisWebin riittämättömyys sen nykyisellä tasolla. Merkittävin ongelma on se, että järjestelmä ei huomioi erävastauskatoa. Siten esimerkiksi yhdellä laskentayksiköllä varustetun junan mittauksia pidetään koko lähtöä edustavina tuloksina, vaikka se olisi todellisuudessa sisältä-

Vastauskadon oikaisumenetelmä	Analyysityökalu	Raportointityökalu
Painotusmenetelmät	DavisWeb	DavisWeb
Painotusmenetelmät	SAS	Ei valmiina
Imputointi	SAS	Ei valmiina

Taulukko 3: *Tilastointiprosessin vaihtoehdot.*

nytkin useamman yksikön. Tämän vuoksi DavisWebin käyttö sellaisenaan tuottaa aivan liian pieniä estimaatteja.

Tilastollisen estimoinnin lisäksi analyysityökalussa tulisi voida tehdä tilastollista editointia. Eräs mittausvirhettä ilmentävä muuttuja on kokonaisnousijoiden ja -poistujien ero, joka ideaalitilanteessa on nolla. DavisWebissä aineistoa voi suodattaa jättämällä pois yksikkömittaukset, joissa nousijoiden ja poistujien perusteella laskettu laatukerroin (kts. kappale (4.2.1)) ylittää jonkin ennalta määrätyn rajan. Laatukertoimen lisäksi Dilax lukitsee dataa, jota voidaan eri tekijöistä johtuen pitää epäluotettavana. Syitä lukitsemiseen ovat esimerkiksi sähkökatkos, järjestelmävirheet ja mahdoton ajoneuvonopeus. Lukittu data on automaattisesti suljettu pois käytöstä, mutta sitä voi halutessaan vapauttaa raportointia varten. Esimerkiksi mahdottoman nopeuden syy voi olla väärin mitatut asemien etäisyydet, joilla ei ole merkitystä mitattujen matkustajamäärien luotettavuuteen [25, 2012]. Editointimahdollisuudet ovat paljon monipuolisemmat SAS-ohjelmistossa, sillä DavisWebin editointityökalujen tueksi tai niitä korvaamaan voidaan kehittää uusia mittausvirheiden indikaattoreita.

Koska laskentalaitteellisen kaluston kierrätystä ei voida suunnitella, kaikkia uniikkeja lähtöjä ei saada mitatuksi joka kuukausi. Taulukkoon (4) on havainnollistettu tammikuun 2013 mittausten aikataulullinen kattavuus eri linjoilla. Ainoastaan A-junasta oli saatu jokaisesta uniikista lähdöstä näyte. Uniikki lähtö katsotaan mitatuksi, mikäli siitä on mitattu vähintään yksi yksikkö. DavisWebin painoa (4.2) parempi vaihtoehto reagoida mittaamattomiin uniikkeihin lähtöihin on jälkiosittaa ne jonkun mahdollisimman homogeenisen mitatun lähdön kanssa samaan soluun. Valitsemalla analyysityökaluksi SAS jälkiositteet voidaan määrittää halutunlaisiksi.

Painotusmenetelmien myötä menetettäisiin mahdollisuus käyttää aiemmilta kuukausilta kerättyä informaatiota. Uniikki lähtö ajetaan toistuvasti viikosta toiseen, eikä lyhyellä aikavälillä matkustajien käyttäytyminen muutu merkittävästi. Poikkeuksena ovat kuitenkin touko- ja kesäkuun sekä elo- ja syyskuun vaihde, jolloin työ- ja koulumatkojen osuus muuttuu merkittävästi. Jos uniikista lähdöstä ei ole mittausta tutkittavalta kuukaudelta, voidaan imputoinnissa etsiä luovuttaja kuukaudelta, jota voidaan pitää matkustajien käyttäytymisen kannalta samankaltaisena. Painotusmenetelmissä mittaamaton uniikki lähtö jouduttaisiin jälkiosittamaan mahdollisimman homogeenisen mitatun lähdön kanssa samaan soluun. Usein tämä tarkoittaa kellonajallisesti lähellä olevien lähtöjen ryhmittämistä yhteen. Peräkkäisten vuorojen matkustajamäärät saattavat kuitenkin poiketa toisistaan paljon, mikä johtuu liityntäliikenteen vaikutuksista. Siksi saman vuoron arvojen imputoin-

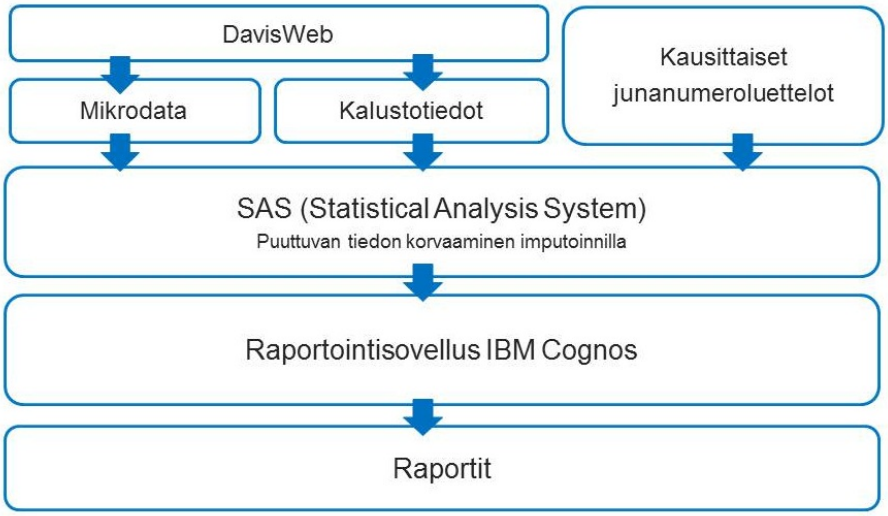
Linja	Mitatut uniikit lähdöt (kpl)	Suunnitellut uniikit lähdöt (kpl)	Kattavuus
A	175	175	100,00 %
E	63	65	96,92 %
H	10	37	27,03 %
I	56	63	88,89 %
K	54	55	98,18 %
L	19	35	54,29 %
M	292	296	98,65 %
N	245	248	98,79 %
R	0	38	0,00 %
S	31	32	96,88 %
T	11	14	78,57 %
U	40	43	93,02 %
Y	12	13	92,31 %
Z	0	40	0,00 %

Taulukko 4: *APC-datan aikataulullinen kattavuus tammikuussa 2013.*

tia aiemmilta ajanjaksoilta eli *last value carried forward*-menetelmää pidetään parempana vaihtoehtona.

Imputoinnin etu on myös raportoinnin helppous. Tulosaineisto on "täydellinen", joten siitä on helppo laskea tunnuslukuja erilaisten taustamuuttujien mukaan. Aineiston käyttö ei vaadi painojen käytön tuntemusta, vaan käyttäjä selviytyy yksinkertaisilla summa- ja jakolaskuilla. Imputointi tosin jättää käyttäjälle vastuun arvioida estimaattien luotettavuutta tarkkailemalla mitattujen (uniikkien) lähtöjen lukumäärää suhteessa mittaamattomiin.

Matkustajamäärätutkimuksen suunnitteluvaiheen työryhmä oli hyvin yksimielinen siitä, että vastauskadon oikaisumenetelmänä tultaisiin käyttämään imputointia. Raportointisovellukseksi valittiin IBM Cognos, joka oli jo ennestään käytössä VR:llä. Lähijunaliikenteen tilastoinnin prosessikuvaus on esitetty kuvassa (2). Mikrodata pitää sisällään APC-laitteiden keräämät asemamittaukset nousijoille ja poistujille sekä tietenkin tarvittavat apumuuttujat ja yksilöintitunnukset. Tieto siitä, mikä yksikkö on ollut mittaamassa mitäkin lähtöä, tulostetaan erillisenä raporttina. Aineistot tuodaan SAS-ohjelmistoon, jossa ne yhdistetään ensin toistensa kanssa ja sen jälkeen voimassa olevien junanumeroluetteloiden kanssa. Tässä vaiheessa saadaan liitettyä suunnitellut junayksikkömäärät kuhunkin



Kuva 2: *Lähijunaliikenteen tilastoinnin prosessikuvaus.*

lähtöön. Tilastollisen editoinnin sekä puuttuvan tiedon imputoinnin jälkeen tulosaineisto viedään Cognokseen, joka automaattisesti osaa tuottaa raportteja käyttäjän asettamien valintojen mukaan.

4.2 Tilastollinen editointi ja imputointi

Datan editointia tarvitaan puhdistamaan aineistoa erilaisista virhemuodoista, kuten laite- tai asennusvikojen aiheuttamista mittausvirheistä, kehikkovirheistä sekä Dilaxin raakadatan työstämisessä tapahtuvista virheistä. Tämän lisäksi tarvitaan laitteiden laskentatarkkuuden seurantajärjestelmä, jonka avulla voidaan löytää systemaattiset laiteviat. Koska analyysityökaluksi valittiin SAS, datan virhesuodatusjärjestelmään voidaan sisällyttää DavisWebin automaattisia suodattimia sekä omia SAS-kielisiä suodatinkoodeja.

Imputoinnilla on VR:n lähijunaliikenteessä tarkoitus kompensoida yksikkövastauskatoa eli mittaamattomia lähtöjä sekä virheelliseksi tai epäluotettavaksi todettuja mittauksia. Imputoinnin tavoite on saada aggregaattitason estimaatit mahdollisimman lähelle oikeita arvoja. Tulosaineiston käyttö rajoittuu matkustajien kokonaismäärien ja keskiarvojen estimointiin, joten jakaumatason onnistumista ei pidetä tärkeänä.

4.2.1 Tilastollinen editointi DavisWebissä

Dilax suorittaa datan editointia lukitsemalla mittauksia, jotka ovat puutteellisia tai tulkittavissa virheellisiksi. Mittausvirheiden löytämiseksi Dilax käyttää tekijöitä kuten, järjestelmäviat, sähkökatkokset, suuri ero kokonaisnousijoiden ja -poistujien välillä, liian suuri ajoneuvonopeus sekä aikataulun liiallinen poikkeavuus suunnitellusta. Puutteelliseksi Di-

lax määrittelee mittaukset, joissa kuorma ei ole tiedossa ensimmäisellä asemalla. [2, 2012.] Näiden lisäksi mittauksia voidaan lukita tapauskohtaisesti, mikäli niiden luotettavuutta on syytä epäillä joistakin muista tekijöistä johtuen. Lukitsemisen voi tehdä ainoastaan tietokannan ylläpitäjä eli Dilax. On huomattava, että Dilaxin suorittama aineiston editointi ei ole lopullista, vaan käyttäjä voi halutessaan vapauttaa Dilaxin lukitsemaa dataa raporttien käyttöön.

Suuret erot kokonaisnousijoiden ja -poistujien välillä viittaavat usein infrapunasensoreiden huonoihin säätöasetuksiin, jolloin esimerkiksi lapset tai kumarassa kyytiin nousevat matkustajat etenkin portaallisissa oviaukoissa eivät tule lasketuksi. Datan editointivaiheessa Dilax laskee jokaiselle mitatulle junayksikölle laatukertoimen (*Quality Level*), joka perustuu kokonaisnousijoiden ja -poistujien prosentuaaliseen eroon. Kun junan henkilökunta jätetään huomioimatta, ideaali tilanne syntyy silloin, kun nousijat ja poistujat ovat yhtä suuret. Mittausvirheiden takia ne eivät aina kuitenkaan vastaa toisiaan. Olkoon A lähdön kokonaisnousijat ja B vastaavasti kokonaispoistujat. DavisWebin laatukerroin QL lasketaan kuten

$$\mathrm{QL} = \begin{cases} 100 \cdot |A - B|/(A + B), & \text{kun } A + B \geq 50 \\ 2 \cdot |A - B|, & \text{kun } A + B < 50, \end{cases}$$

jolloin se vaihtelee välillä $0 \leq \mathrm{QL} \leq 100$ [3, 2012]. DavisWebin *Filter*-alueella laatukertoimelle voi asettaa maksimin, jolloin raportissa jätetään huomioimatta rajan ylittävät mittaukset. Dilaxin mukaan laatukerroin 10 on vielä hyväksyttävä, mutta jos laskentayksikkö systemaattisesti mittaa sitä suurempia laatukertoimia, on kyseessä usein tekninen ongelma [25, 2012]. Mikäli laskentayksikössä ilmenee systemaattinen vika, raportoinnista voidaan sulkea kokonaisia yksiköitä pois.

Laatukerroin ja datan lukitsemisperiaatteet palvelevat hyvin editointitarpeita, joten niitä tullaan käyttämään lähijunaliikenteen matkustajamäärätutkimuksen ensimmäisen asteen editoinnissa (kts. kuva (4)). Toisen asteen editointi tapahtuu SAS-ohjelmistossa, jonne kehitetään muihin virhelähteen muotoihin perustuvia suodattimia.

4.2.2 Tilastollinen editointi SAS-ohjelmistossa

Peruutettujen lähtöjen muodostamasta ylipeitosta ja sen kärsimisestä puhuttiin jo kappaleessa (2.2.3). Muista tekijöistä johtuva yli- ja alipeitto ovat esiintyessään havaittavissa mikrodatan ja junanumeroluettelon yhdistämisvaiheessa. Yli- tai alipeitoksi voidaan tulkita sellaiset lähdöt, joille ei löydy vastinetta eli paria toisesta aineistosta. Mikäli vastinetta ei löydy junanumeroluettelosta, on kyseessä ylipeitto, ja vastaavasti parin uupuessa mik-

Editointivaihe	Ohjelmisto	Virhemuoto	Työkalu
1. asteen editointi	DavisWeb	Nousijoiden ja poistujien absoluuttinen erotus	Laatukerroin, QL (Max 10)
		Järjestelmäviat Sähkökatkokset Aikataulun vaihtelevuus Tuntematon kuorma ensimmäisellä asemalla Mahdoton nopeus Muut yksittäiset lukitsemisen syyt	Datan lukitseminen
		Laskentayksikössä havaittu systemaattinen vika	Suodatin laskentayksikön yksilöintitunnuksen perusteella*
2. asteen editointi	SAS-ohjelmisto	Muusta kuin peruutetuista lähdöistä muodostuva yli- ja alipeitto	MERGE- komennolla
		Sm2- ja Sm5-yksikkö samalla lähdöllä	DATA step- komennolla
		Kokonaismatkustajamäärän 50 %:n poikkeavuus saman vuoron keskiarvosta**	DATA step- komennolla
		Erävastauskato	Keskiarvoimputointi

*Käytetään tarvittaessa
**Näytteitä ei poisteta aineistoista

Kuva 3: *Editointivaiheet ja -periaatteet matkustajamäärätutkimuksessa.*

rodatasta on kyse alipeitosta.

Dilaxin onnistumista raakadatan kohdistamisessa oikeille linjoille ja lähdöille on vaikea seurata hallitusti. Kaluston perusteella on kuitenkin mahdollista karsia deduktiivisesti pois lähdöt, jotka ovat todellisuudessa mahdottomia. Sm2- ja Sm5-junayksikköjä ei ole mahdollista kytkeä toisiinsa. Toisinaan mikrodatassa esiintyy lähtöjä, joissa mittaavana yksikkönä on ollut sekä Sm2- että Sm5-kalustosarjaa oleva yksikkö. Tällöin datan kohdistamisessa on täytynyt tapahtua virhe, joten mittaukset poistetaan aineistosta DATA step-komennolla.

Systemaattisten laitevikojen havaitsemiseksi lähdöt, joissa kokonaismatkustajamäärä eroaa vähintään 50 % vuoron keskimääräisestä tasosta, merkitään seurantaa varten. Mikäli jokin yksikkö toistuvasti tuottaa poikkeavia matkustajamääriä, voidaan alkaa epäillä systemaattista laitevikaa. Näytteitä ei kuitenkaan poisteta aineistosta, koska esimerkik-

si poikkeustapahtumat ja -ajankohdat, kuten esimerkiksi konsertit, urheilutapahtumat tai vappu, saattavat hetkellisesti nostattaa vuoron matkustajamäärää. Näihin voidaan toisaalta varautua, mutta esimerkiksi junan muutaman minuutin myöhästymisellä saattaa olla huomattavia vaikutuksia liityntäliikenteeseen. Yksittäisen lähdön kokonaismatkustajamäärästä tai maksimikuormasta on siis vaikea tehdä virhepäätelmiä. Ainoastaan junahenkilökunnan antamat lausunnot tai kaluston kapasiteetin reilusti ylittävä maksimikuorma voivat todistaa yksittäisen matkan mittaukset virheellisiksi.

Erävastauskatoa eli vajavaisia lähtöjä kompensoidaan käyttämällä keskiarvoimputointia. Yksiköiden sijaintia toisiinsa nähden ei tiedetä, mutta järjestyksen tiedetään olevan likimain satunnainen. Olkoon eräällä lähdöllä T kappaletta pysähtymisasemia. Olkoon lähdöllä i suunniteltuja yksikköjä i_s kappaletta, joista laskentatulokset on saatu i_m, $1 \leq i_m < i_s$, kappaleesta yksiköitä. Olkoon $a_k(m)$ yksikön k mittaamat poistujat asemalla m ja vastaavasti $b_k(m)$ yksikön k mittaamat nousijat asemalla m. Tällöin mittaamattomille yksiköille lainataan keskiarvot

$$\begin{aligned} \overline{a}_{obs}(m) &= \left(\sum_{k=1}^{i_m} a_k(m) \right) / i_m, \quad m = 1, \ldots, T, \\ \overline{b}_{obs}(m) &= \left(\sum_{k=1}^{i_m} b_k(m) \right) / i_m, \quad m = 1, \ldots, T. \end{aligned}$$

4.2.3 Imputointimenetelmät

DavisWebistä tulostettu mikrodata sisältää apumuuttujat lähtöaika, päivämäärä, viikonpäivä, linja ja suunta. Myöhemmin yhdistettäessä mikroaineisto voimassa olevien junanumeroluetteloiden sekä kalustotietojen kanssa saadaan tieto lähdön yksikkömäärästä, junanumerosta sekä mittaavista yksiköistä. Mikrodataan tulostetaan aina koko kehikkoperusjoukko eli myös mittaamattomat lähdöt.

Aineisto jaetaan imputointisoluihin linjan, päivätyypin, suunnan ja lähtöajan mukaan. Päivät maanantaista torstaihin luokitellaan yhdeksi päivätyypiksi, koska matkustajien käyttäytyminen on silloin melko samanlaista. Loput päivät eli perjantai, lauantai ja sunnuntai jaotellaan erillisiksi päivätyypeiksi. Koska tulosaineiston käyttötarkoitus on melko yksinkertainen rajautuen kokonaismäärien ja keskiarvojen estimointiin, päätettiin imputointisolun sisällä käyttää mediaani-imputointia ilman satunnaistermiä. Mediaanin käyttäminen perustuu siihen, että tavallisesta poikkeavat ääripään mittaukset eivät pääse vaikuttamaan imputointiarvoon niin paljon kuin keskiarvossa. Tällaisia poikkeavia mittauksia voi aiheuttaa esimerkiksi peruuntunut juna, jolla voi olla merkittäviä vaikutuksia seuraavan vuoron matkustajamäärään.

Kaikista imputointisoluista ei välttämättä saada näytteitä tutkittavan kuukauden aikana. Tyhjien solujen osalta turvaudutaan *last value carried forward*-menetelmään, jossa arvot imputoidaan matkustajakäyttäytymisen kannalta samanlaiselta kuukaudelta. Toimintaperiaate on samanlainen kuin edellä kuvattu lainaussääntö epätyhjille imputointisoluille; ainoa poikkeus on lainauskuukauden vaihtaminen edeltävään ajanjaksoon. Luovuttavana imputointisoluna toimii siis samat linjaan, päivätyyppiin, suuntaan ja lähtöaikaan sidotut lähdöt, joiden mediaani lainataan. Jos tarkastelukuukaudelle ei löydy tarpeeksi samankaltaista kuukautta, voidaan kausivaihtelu ottaa huomioon kertomalla arvoja kausikertoimella. On hyvä huomioida, että lainaus suoritetaan aina mitatuilta lähdöiltä, eli lainattua arvoa ei koskaan lainata eteenpäin.

Mikäli imputointisolusta ei ole olemassa näytteitä tutkimuskuukaudelta eikä edeltäviltä kuukausilta, luovuttaja etsitään käyttämällä lähtöaikaa läheisyyden mittana saman linjan, suunnan ja päivätyypin sisällä. Kun kellonajallisesti lähin mitattu uniikki lähtö on löytynyt, imputoitavana arvona käytetään uniikista lähdöstä saatujen mittausten mediaania. Lainatut arvot voivat olla joko tutkittavalta kuukaudelta tai aiemmalta ajanjaksolta riippuen siitä kummasta aineistosta lähin mitattu uniikki lähtö löytyy.

Imputoinnista jätetään aineistoon jälki, jotta tulosaineiston käyttäjä osaa erottaa todelliset ja lainatut arvot. Jälki sisältää myös tiedon siitä, mitä kolmesta edellä kuvatusta menetelmästä lainauksessa on käytetty.

Poikkeuspäivät, kuten arkipyhät ja muut lakisääteiset juhlapäivät, poikkeavat liikenteeltään tavanomaisesta päivästä. Tämä haluttiin huomioida imputoinnissa siten, että poikkeuspäivien mittaukset asetettiin edustamaan vain itseään. Niitä ei siis käytetä imputoinnissa, mutta lainaus poikkeuspäiville suoritetaan kuitenkin kuten muillekin. Poik-

Imputointisolusta näytteitä tutkittavalta kuukaudelta?	Imputointisolusta näytteitä edellisiltä kuukausilta?	Imputointiperiaate	Lainauskuukausi
Kyllä	–	Mediaani-imputointi imputointisolun sisällä	Tutkimuskuukausi
Ei	Kyllä	Mediaani-imputointi imputointisolun sisällä	Aikaisempi kuukausi
Ei	Ei	Lähtöaika läheisyyden mittana	Tutkimuskuukausi / aikaisempi kuukausi

Kuva 4: *Imputointiperiaatteet.*

keuspäivän liikennöinti toteutetaan usein poiketen tavanomaisesta kalenteriviikonpäivästä, joten päivätyyppi vaihdetaan tarpeen mukaan, ja tämä otetaan huomioon imputoinnissa.

Mittausvirheiden takia lähdön kokonaisnousijat ja -poistujat eivät aina ole yhtäsuuret. Raportointia varten ne kuitenkin tasataan. Kullekin lähdölle i lasketaan prosentuaalinen kokonaispoistujien A_i ja -nousijoiden B_i ero

$$p_i = (A_i - B_i) / (A_i + B_i), \tag{4.4}$$

jonka verran nousijat ja -poistujat pakotetaan tulemaan toisiaan vastaan. Lähdöllä i olkoon $a_i(m)$ poistujat ja $b_i(m)$ nousijat asemalla m. Korjauskertoimella (4.4) saatavat uudet arvot ovat tällöin

$$\begin{aligned} a_i^*(m) &= (1 + p_i)a_i(m) \\ b_i^*(m) &= (1 - p_i)b_i(m). \end{aligned}$$

Nyt kokonaisnousijat ja -poistujat eli muuttujien $b_i^*(m)$ ja $a_i^*(m)$ summat yli asemien m ovat samat:

$$A_i^* = \sum_m a_i^*(m) = \sum_m b_i^*(m) = B_i^*. \tag{4.5}$$

Luvut $a_i^*(m)$ ja $b_i^*(m)$ eivät välttämättä ole kokonaislukuja, joita raporteissa pitäisi ainoastaan esiintyä. Pyöristämällä $a_i^*(m)$ ja $b_i^*(m)$ lähimpään kokonaislukuun ei kaava (4.5) välttämättä enää päde. Tästä syystä pyöristetyillä luvuilla täytyy tehdä toinen tasaus. Tasaus suoritetaan aina Helsingin päässä, koska siellä matkustajien liikkuvuus on suurinta.

Olkoon $b_i^1(m)$ ja $a_i^1(m)$ lähdön i korjauskertoimella (4.4) korjatut ja lähimpään kokonaislukuun pyöristetyt nousija- ja poistujamäärät asemalla m. Merkitään lähdön i kokonaisnousijoiden ja -poistujien erotusta symbolilla $\varepsilon(i)$. Kun suunta on Helsingistä, toinen tasaus suoritetaan kuten

$$\begin{aligned} b_i^2(HKI) &= b_i^1(HKI) - \varepsilon(i) \\ a_i^2(HKI) &= a_i^1(HKI). \end{aligned}$$

Muilla asemilla, eli kun $m \neq HKI$, valitaan $b_i^2(m) = b_i^1(m)$ ja $a_i^2(m) = a_i^1(m)$. Kun junan suunta on kohti Helsinkiä, valitaan

$$\begin{aligned} b_i^2(HKI) &= b_i^1(HKI) \\ a_i^2(HKI) &= a_i^1(HKI) + \varepsilon(i). \end{aligned}$$

ja kaikille $m \neq HKI$ valitaan $b_i^2(m) = b_i^1(m)$ ja $a_i^2(m) = a_i^1(m)$.

4.3 Raportointi

Raportointijärjestelmäksi valittiin IBM Cognos, jonka käyttöliittymä on web-pohjainen. Työasemiin ei tarvitse tehdä erillisiä ohjelmistoasennuksia, mikä helpottaa järjestelmän käyttöönottoa ja minimoi ylläpito- ja asennuskustannukset. Ohjelmisto on ollut aiemmin käytössä VR:llä, joten he vastaavat ohjelmiston ylläpidosta sekä käyttäjäryhmien ja -oikeuksien hallinnasta.

Järjestelmästä on saatavilla neljä vakioraporttia: liikennepaikkojen matkustajamäärät (raportti 1), linjakohtainen matkustajamäärä (raportti 2), junakohtainen kokonaiskäyttö (raportti 3) sekä täyttöaste (raportti 4). Tarkemmat raporttikuvaukset on esitetty myöhempänä. Raportit jäljittelevät manuaalilaskennoista tehtyjä ns. perusraportteja. Koska vanhat manuaalilaskennoista tehdyt raportit edustivat vain yhden arkipäivän, lauantain ja sunnuntain aikana ajettuja lähtöjä, on uudessa raportoinnissa otettu huomioon mahdollisuus tuottaa summaraporttien lisäksi keskimääräistä päivää kuvaava raportti. Raporteissa 1, 2 ja 3 on valittavana kolme aggregaattifunktiota: summa, keskiarvo ja mediaani. Summa tuottaa valittujen päivien summan, keskiarvo päiväkohtaisen keskiarvon ja mediaani päiväkohtaisen mediaanin. Jokaisessa raportissa mittausaste eli mitattujen yksikköjen suhde kaikkiin yksiköihin voidaan valita joko näkyväksi tai ei-näkyväksi.

Muokattavia raportteja varten on olemassa erityinen raportointikuutio, jossa käyttäjä voi itse luoda oman raportin. Kuutiossa käyttäjä voi määritellä aineiston suodattimet sekä valita taustamuuttujat ja niille esitettävät mittarit. Kuutio toimii vakioraportoinnin tukena tuottamaan suunnittelua tukevia tai erilaisiin projekteihin tai tutkimuksiin tarvittavia raportteja. Sekä vakioraportit että raportointikuutiolla tehdyt raportit voi tulostaa esimerkiksi Excel-, PDF- ja HTML-muodossa.

Liikennepaikkojen matkustajamäärät

Liikennepaikkojen matkustajamäärät-raportissa tarkastellaan asemakohtaisia nousijoita, poistujia ja niiden summaa eri rataosuuksilla. Rataosuudet ovat rantarata, Vantaankosken rata, päärata sekä oikorata. Käyttäjä voi itse suodattaa raportissa käytettävää aineistoa valitsemalla viikonpäiväryhmän, aikajakson ja suunnan.

Raportin alalaitaan tulostuu yhteenveto, josta nähdään rataosuuden kokonaiskäyttö eriteltynä vielä YTV-, HSL- ja vyöhykealueisiin. Raportti on tärkeä erityisesti matkustajien kokonaismäärän ja sen kehityksen seurannassa.

Linjakohtainen matkustajamäärä

Linjakohtainen matkustajamäärä-raportti on hyvin samanlainen kuin edellinen, mutta rataosuuden sijasta tuloksia tarkastellaan linjakohtaisesti. Nousijoiden ja poistujien lisäksi raportissa on mittarina myös kuorma. Aineistoa voi suodattaa viikonpäiväryhmän, aikajakson, rataosan ja suunnan mukaan.

Kuormaan on aina syytä liittää yhteysväli eli junan lähtö- ja pääteasema, joiden väliltä kuorma on mitattu. Raportissa kuorma on kuitenkin esitetty asemakohtaisesti, joten junan kulkusuunta on otettava huomioon raporttia lukiessa. Aseman kohdalla esitetty kuorma kuvaa aina kyseiseltä asemalta lähteneiden junien kuormaa. Mikäli suodattimeen on valittu molemmat suunnat (sekä Helsinkiin että Helsingistä), on huomioitava, että kuorman laskennassa on käytetty kahta eri yhteysväliä.

Raportin alalaidassa on yhteenveto, josta nähdään linjakohtainen kokonaiskäyttö eriteltynä YTV-, HSL- ja vyöhykealueisiin. Tällä on merkitystä esimerkiksi HSL:n VR:lle maksamien lipputulomenetyskorvausten laskennassa YTV-alueen ulkopuolella.

Junakohtainen kokonaiskäyttö

Junakohtainen kokonaiskäyttö-raportissa esitetään yksittäisen junanumeron alla kulkevien lähtöjen asemakohtaisia nousijoita, poistujia, kuormia sekä täyttösuhteita. Täyttösuhde tarkoittaa kuorman suhdetta kaluston kapasiteettiin eli istuma- ja seisomapaikkojen lukumäärään. Aineistoa voi suodattaa viikonpäiväryhmän, aikajakson, kellonajan, suunnan ja linjan perusteella.

Raporttia käytetään esimerkiksi kapasiteetintarpeen suunnittelussa. Mikäli täyttösuhde on tietyllä lähdöllä jatkuvasti korkea, on siihen syytä lisätä yksi yksikkö vastaamaan paremmin kysyntää. Käänteisesti lähdöltä voidaan vähentää yksikkö, jos täyttösuhde on jatkuvasti alhainen ja lähdöllä on alunperin ollut vähintään kaksi yksikköä.

Raportin alalaidassa on yhteenveto, josta nähdään junanumerokohtainen kokonaiskäyttö eriteltynä YTV-, HSL- ja vyöhykealueisiin.

Täyttöaste

Täyttöaste-raportti etsii jokaiselta junanumerolta täyttöasteeltaan suurimman yhteysvälin ja sen kuorman. Raporttiin tulostuu kaksi taulukkoa, joista toisessa esitetään junanumerolle täyttöasteeltaan keskimääräisesti suurin asemaväli. Toisessa taulukossa esitetään junanumeron täyttöasteeltaan suurin yksittäinen yhteysväli.

Taulukon lajitteluperusteen voi valita junan (nouseva), lähtöajan (nouseva), kuorman

(laskeva) tai täyttösuhteen (laskeva) perusteella. Aineistoa voi suodattaa viikonpäiväryhmän, aikajakson, kellonajan, suunnan ja linjan perusteella.

Täyttöaste-raportilla voidaan tarkastella kunkin lähdön vilkkaimpia asemavälejä. Raporttia käytetään esimerkiksi kapasiteetintarpeen suunnittelun apuna.

5 Markovin teoriaa

5.1 Stokastiset prosessit

Stokastiset prosessit kuvaavat matemaattisesti ajassa eteneviä todellisuuden prosesseja, joissa siirtyminen tilasta toiseen on satunnaista. Tällaisia prosesseja esiintyy runsaasti luonnossa ja siksi stokastisten prosessien teorialla on merkittäviä sovellusalueita eri tieteen aloilla, kuten lääketieteessä, taloustieteessä, biologiassa ja fysiikassa. Tyypillisiä luonnossa ilmeneviä stokastisia prosesseja ovat populaation kannan vaihtelu, taudin leviäminen, asiakaspalvelupisteen asiakasvirta ja yksittäisen molekyylin muodostama Brownin liike. Edellä kuvattujen ilmiöiden kehitystä ei voida etukäteen tietää, mutta stokastisten prosessien teorian avulla niitä voidaan mallintaa, ja siten laatia ennusteita tulevaisuudesta.

Matemaattisesti stokastinen prosessi $\{X_t \mid t \in I\}$ on perhe satunnaismuuttujia X_t, joiden alkeistapakset on koottu tilajoukkkoon S. Indeksi t on aikaparametri, ja sen arvot on koottu indeksijoukkoon I, joka voi olla joko jatkuva tai diskreetti. Tässä tutkielmassa tullaan jatkossa käsittelemään vain diskreettiaikaisia stokastisia prosesseja, joissa tilajoukko on joko luonnolliset luvut $\mathbb{N}$ tai äärellinen joukko $\{0, 1, 2, \ldots, n\}$.

Ennen stokastisten ja erityisesti Markovin prosessien ominaisuuksien tarkempaa esittelyä täytyy luoda mittateoreettinen pohja, joka tekee tulevista todennäköisyyteen liittyvistä tarkasteluista hyvin määriteltyjä. Satunnaisilmiön tapahtumat käsitetään mitallisina joukkoina, joiden mitta vastaa tapahtuman todennäköisyyttä. Mittateoreettisen määrittelyn huipentuma tulee olemaan tässä satunnaismuuttujan määritelmä, jossa kiteytyy kaikki todennäköisyysteorian tärkeimmät käsitteet.

5.2 Satunnaismuuttuja ja todennäköisyysavaruus

Todennäköisyyslaskennassa ollaan kiinnostuneita satunnaisilmiössä havaittavissa olevien tai muuten mielenkiintoisten tapahtumien todennäköisyyksistä. Satunnaismuuttuja X on havaittavissa olevan muuttujan numeerinen vastine, jonka todennäköisyysjakauma koostuu muuttujan arvoista ja niihin liittyvistä todennäköisyyksistä. Jotta satunnaismuuttujan määritelmä voidaan formaalisti esittää, täytyy ensin määritellä todennäköisyysavaruus.

Otosavaruus (*sample space*) Ω on kaikkien mahdollisten satunnaisilmiön tulosten joukko. Jos otosavaruus on numeroituva tai numeroituvasti ääretön, sitä sanotaan diskreetiksi, muulloin se on jatkuva. Merkitään otosavaruuden Ω potenssijoukkoa eli osajoukkojen perhettä symbolilla $\mathcal{B}(\Omega)$. Potenssijoukon alkiot ovat tapahtumia, jotka voivat esiintyä satunnaisilmiössä, mutta joille ei välttämättä voida liittää todennäköisyyttä johdonmukaisella tavalla. Tällöin on tyydyttävä potenssijoukon osajoukkoihin. [22, s.9, 2006.]

Määritelmä 5.1. Sigma-algebra $\mathcal{F}$ on perhe otosavaruuden Ω osajoukkoja siten, että seuraavat aksioomat ovat voimassa:

(i) Otosavaruudelle pätee $\Omega \in \mathcal{F}$.

(ii) Jos tapahtumalle A pätee $A \in \mathcal{F}$, niin tapahtuman A komplementille pätee $A^c \in \mathcal{F}$.

(iii) Jos on voimassa $A_k \in \mathcal{F}$ kaikilla $k \in K$, missä K on numeroituva joukko, niin tapahtumien A_k yhdisteelle pätee

$$\bigcup_{k \in K} A_k \in \mathcal{F}.$$

Yleensä riittää valita pienin sigma-algebra, joka sisältää kiinnostuksen kohteena olevat tapahtumat [20, s.9, 1999]. Äärellisen otosavaruuden ollessa kyseessä sigma-algebraksi valitaan kuitenkin lähes aina otosavaruuden potenssijoukko $\mathcal{B}(\Omega)$, joka toteutaa triviaalisti määritelmän (5.1) ehdot (i)-(iii). Tällöin todennäköisyys on määritelty jokaiselle otosavaruuden osajoukolle.

Todennäköisyysmitta (*probability measure*) liittää jokaiseen tapahtumaan eli sigma-algebran alkioon A luvun $\mathbb{P}(A)$, jota sanotaan tapahtuman A todennäköisyydeksi. Luku $\mathbb{P}(A)$ ilmaisee todennäköisyyden, että satunnaismuuttujan X arvo kuuluu joukkoon A.

Määritelmä 5.2. Todennäköisyysmitta $\mathbb{P}$ on kuvaus $\mathbb{P} : \mathcal{F} \to [0, 1]$, joka toteuttaa seuraavat ehdot:

(i) $\mathbb{P}(\Omega) = 1$.

(ii) Erillisille tapahtumille $A_1, A_2, \ldots \in \mathcal{F}$ pätee

$$\mathbb{P}\{\bigcup_{n=1}^{\infty} A_n\} = \sum_{n=1}^{\infty} \mathbb{P}\left(A_n\right).$$

Todennäköisyysavaruus (*probability space*) on otosavaruuden, tapahtumien osajoukon ja todennäköisyysmitan muodostama kolmikko $(\Omega, \mathcal{F}, \mathbb{P})$, jota toisinaan nimitetään myös

todennäköisyyskentäksi. On muistettava, että satunnaisilmiötä kuvaava todennäköisyysavaruus ei ole yksikäsitteinen, vaan se voidaan määritellä sovelluskohteesta ja tutkimustilanteesta riippuen.

Satunnasmuuttuja määritellään todennäköisyysavaruudessa kuvaukseksi, joka liittää jokaiseen alkeistapaukseen jonkin reaalilukuarvon.

Määritelmä 5.3. Satunnaismuuttuja X on todennäköisyyskentässä $(\Omega, \mathcal{F}, \mathbb{P})$ määritelty $\mathcal{F}$-mitallinen kuvaus $X : \Omega \to \mathbb{R}$. Tämä tarkoittaa sitä, että jokaisella reaalilukujen Borel-joukolla[1] B pätee

$$X^{-1}\{B\} := \{\omega : X(\omega) \in B\} \in \mathcal{F},$$

missä alkeistapakset ω ovat otosavaruuden alkioita.

5.3 Markovin ketju

Markovin prosessi on stokastinen prosessi, jossa tapahtumaketjun tulevaisuuteen vaikuttaa vain nykytilanne. Ennustettaessa tulevia tapahtumia ei nykyhetkeä edeltäneellä historialla ole siten merkitystä. Markovin prosessit voidaan jakaa neljään eri luokkaan niiden tilajoukon ja aikaparametrin indeksijoukon tyypin perusteella. Diskreetin tilajoukon ollessa kyseessä Markovin prosesseja kutsutaan Markovin ketjuiksi, joiden diskreettiaikaiseen teoriaan tullaan jatkossa rajoittumaan. Tutkielmassa kuvataan Markovin ketjut siinä laajuudessa, jota tarvitaan empiirisen prosessin Markov- ominaisuuden tutkimiseen.

5.3.1 Markov-ominaisuus

Stokastisten prosessien yhteydessä otosavaruudesta käytetään nimeä tilajoukko, jota tavataan merkitä kirjaimella S. Oletetaan, että $X_n \in \mathbb{R}^d$. Jos satunnaismuuttuja X_n saa tuloksen x, sanotaan, että stokastinen prosessi on hetkellä n tilassa x. Koska satunnaismuuttujat ovat riippuvaisia nykyhetken tilasta, todennäköisyydet $\mathbb{P}(X_n = x)$ eivät ole vakioita. Tästä syystä on järkevää käsitellä todennäköisyyksiä siirtyä tilasta toiseen.

Jatkossa Markovin ketjut oletetaan aikahomogeenisiksi, eli jokaisella ajanhetkellä $n \geq 0$ Markovin ketju joko pysyy samassa tilassa tai se tekee hypyn johonkin toiseen tilaan [18, s.217, 2002]. Oletetaan, että

$$\begin{aligned} \mathbb{P}\left(X_n \in dy \,|\, X_{n-1} = x\right) &= p(x \to y)dy \quad \text{ja} \\ \pi(A) &= \int_A \pi(x)dx \end{aligned}$$

[1] Reaalilukujen Borelin sigma-algebra $\mathcal{B}$ on suppein reaalilukujen sigma-algebra, joka pitää sisällään avointen joukkojen kokoelman [22, s.11, 2006]. Joukot $B \in \mathcal{B}$ ovat Borel-joukkoja.

ovat tiheyksiä Lebesgue mitan suhteen. Merkitään jatkossa $p(x \to x) = r(x)$. Funktiota $p(x \to y)$ sanotaan alitodennäköisyystiheysfunktioksi (*subprobability density distribution*), jolle pätee

$$\int p(x \to y) \leq 1.$$

Tällöin todennäköisyys, että Markovin ketju tekee hypyn, on

$$\int p(x \to y)dy = 1 - r(x).$$

Määritelmä 5.4. Todennäköisyysavaruudessa $(S, \mathcal{F}, \mathbb{P})$ määritelty stokastinen prosessi $\{X_t \mid t \in \mathbb{N}\}$ on Markovin ketju, mikäli se toteuttaa Markov- ominaisuuden

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_0 = x_0, \ldots X_n = x_n) \quad = \quad \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

kaikilla ajanhetkillä $n \in \mathbb{N}$.

Markov-ominaisuutta sanotaan kuvainnollisesti myös unohtavaisuus-ominaisuudeksi.

5.3.2 Siirtymäydin

Määritelmä 5.5. Olkoon $(X, \mathcal{X})$ ja $(Y, \mathcal{Y})$ mitallisia avaruuksia. Siirtymäydin (*transition kernel*) avaruudesta $(X, \mathcal{X})$ avaruuteen $(Y, \mathcal{Y})$ on funktio $P : X \times Y \to [0, \infty)$, jolle pätee seuraavat ominaisuudet:

(i) kaikilla $x \in X, P(x, \cdot)$ on positiivinen

(ii) kaikilla $A \in \mathcal{Y}$, funktio $x \mapsto P(x, A)$ on mitallinen.

Jos $X = Y$ ja $P(x, X) = 1$ kaikilla $x \in X$, niin P on Markovin siirtymäydin.

Todennäköisyys siirtyä tilasta x joukon A tilaan on

$$\begin{aligned} P(x, A) &= \mathbb{P}(X_{n+1} \in A \,|\, X_n = x) \\ &= \int_A p(x \to y)dy + r(x)\delta_x(A), \quad x \in S, A \subset S, \end{aligned} \tag{5.1}$$

missä $\delta_x(A) := 1$, kun $x \in A$, ja $\delta_x(A) := 0$, kun $x \in A^c$. Todennäköisyyksiä siirtyä tilasta toiseen kutsutaan siirtymätodennäköisyyksiksi (*transition probabilities*). Kun tila-avaruus on äärellinen tai korkeintaan numeroituva, siirtymätodennäköisyydet kerätään usein matriisiksi $\mathbf{P}$, joka on muotoa

$$\mathbf{P} = \begin{bmatrix} r(1) & p(1 \to 2) & \ldots & p(1 \to i) \\ p(2 \to 1) & r(2) & \ldots & p(2 \to i) \\ \vdots & \vdots & \ddots & \vdots \\ p(i \to 1) & p(i \to 2) & \ldots & r(i) \end{bmatrix}$$

5.4 Suurten lukujen laki Markovin ketjulle

Tässä kappaleessa todistetaan suurten lukujen laki (SLL) Markovin ketjulle. Aluksi esitetään määritelmiä ja aputuloksia, jotka ovat hyödyllisiä varsinaista todistusta silmällä pitäen. Tämän kappaleen sekä kappaleen (5.5) lähteenä on käytetty Esa Nummelinin kirjoitusta *MC's for MCMC'ists* lehdestä *International Statistical Review (2002).*

Määritelmä 5.6. Epätyhjää osajoukkoa $I \subset E$ sanotaan pieneksi (*small*), jos on olemassa epätyhjä osajoukko $J \subset E$ ja positiivinen vakio β siten, että

$$p(x \to y) \geq \beta \quad \text{kaikilla } x \in I, y \in J.$$

Määritelmä 5.7. Todennäköisyystiheysfunktio[2] $\pi = (\pi(x))_{x \in E}$ on invariantti, jos

$$\pi \mathbf{P} = \pi$$

Suurten lukujen laki päätee vain seuraavat hypoteesit täyttäville Markovin ketjuille.

Hypoteesi 5.8. On olemassa pieni joukko I siten, että jokaisella alkutilalla $x \in E$ pätee

$$P^{n(x)}(x, I) := \mathbb{P}(X_{n(x)} \in I \,|\, X_0 = x) > 0,$$

missä $n(x) > 1$ on alkutilasta x riippuva aikaindeksi.

Hypoteesi 5.9. (i) Markovin ketjulla on invariantti todennäköisyystiheysfunktio $\pi = (\pi(i))_{i \in E}$.

(ii) invariantin todennäköisyystiheysfunktion π kantajoukko (*support*)

$$S := \{x \in E \,:\, \pi(x) > 0\}$$

toteuttaa ehdon

$$P(x, S) = 1 \quad \text{kaikilla } x \in S.$$

Funktio $f(x)$, $x \in E$, on π-integroituva, jos

$$\pi(f(x)) := \int f(x)\pi(x)dx < \infty.$$

Merkitään funktion f aritmeettista keskiarvoa n:ssä ensimmäisessä tilassa kuten

$$\hat{\pi}_n(f) := n^{-1}(f(X_0) + \ldots f(X_{n-1})).$$

Teoreema 5.10. *Oletetaan, että hypoteesit (1) ja (2) ovat voimassa. Tällöin kaikilla π-integroituvilla funktioilla f ja alkutiloilla $X_0 = x \in S$*

$$\mathbb{P}\left[\lim_{n \to \infty} \widehat{\pi}_n(f) = \pi(f)\right] = 1$$

Lebesgue mitan suhteen.

[2]Funktio $\pi(y)$ on todennäköisyystiheysfunktio, jos $\int \pi(y)dy$=1 on voimassa.

5.4.1 Uusiutuminen

Tässä kappaleessa näytetään, että Markovin ketjun muodostama polku voidaan jakaa riippumattomiin ja samoinjakautuneisiin jaksoihin.

Olkoon ν tasainen todennäköisyystiheysfunktio joukolle J ja nolla tämän komplementille:

$$\nu(y) \equiv \begin{cases} \frac{1}{|J|}, & \text{kun } y \in J \\ 0, & \text{muulloin.} \end{cases}$$

Merkintä $|J|$ tarkoittaa tässä joukon J alkioiden lukumäärää. Määritellään lisäksi funktio

$$s(x) \equiv \begin{cases} \beta|J|, & \text{kun } x \in I \\ 0, & \text{muulloin.} \end{cases}$$

Huomataan, että hypoteesista (5.8) seuraa

$$p(x \to y) \geq s(x)\nu(y) = \begin{cases} \beta, & \text{kun } x \in I \text{ ja } y \in J \\ 0, & \text{muulloin.} \end{cases}$$

Olkoon

$$\begin{aligned} Q(x,A) &:= P(x,A) - s(x)\int_A \nu(y)dy \\ &= \begin{cases} P(x,A) - \beta|J| \cdot |A \cap J| \cdot (|J|)^{-1}, & \text{kun } x \in I, A \subset E \\ P(x,A), & \text{kun } x \in I^c, A \subset E \end{cases} \\ &= \begin{cases} P(x,A) - \beta|A \cap J|, & \text{kun } x \in I, A \subset E \\ P(x,A), & \text{kun } x \in I^c, A \subset E. \end{cases} \end{aligned}$$

Huomataan, että aina pätee $Q(x,A) \geq 0$. Kun $x \in I^c$, epäyhtälö on triviaali. Kun $x \in I$, niin

$$\begin{aligned} Q(x,A) &= P(x,A) - \beta|J \cap A| \\ &\geq \beta|A| - \beta|J \cap A| \\ &\geq 0. \end{aligned}$$

Olkoon satunnaismuuttujalla X_n alitodennäköisyystiheysfunktio $\lambda(x)$ annettu siten, että

$$Q(X_n \in A) = \int_A \lambda(x)dx.$$

Tällöin seuraavan tilan X_{n+1} alitodennäköisyystiheysfunktio on muotoa

$$\lambda Q(y) := \lambda P(y) - \nu(y)\int \lambda(x)s(x)dx,$$

koska

$$\begin{aligned} Q(X_{n+1} \in A) &= \int \lambda(x) Q(x, A) dx \\ &= \int \int_A \lambda(x) p(x, y) dx dy + \int_A \lambda(x) r(x) dx \\ &\quad - \int \lambda(x) s(x) \int_A \nu(y) dx dy \\ &= \int_A \lambda P(y) - \int_A \nu(y) \left[\int \lambda(x) s(x) dx \right] dy \\ &= \int_A \left[\lambda P(y) - \nu(y) \int \lambda(x) s(x) dx \right] dy. \end{aligned} \tag{5.2}$$

Määritellään kaksiulotteinen Markovin ketju (X_n, Y_n) siten, että satunnaismuuttujat $X_0, X_1, \ldots$ ovat tilasekvenssin E alkioita ja satunnaismuuttujat $Y_0, Y_1, \ldots$ ovat binäärisen joukon $\{0, 1\}$ alkioita. Siirtymätodennäköisyydet määritellään kuten

$$\mathbb{P}(X_n \in A\,;\, Y_n = 1 \,|\, X_{n-1} = x\,;\, Y_{n-1}) := s(x) \int_A \nu(y) dy, \tag{5.3}$$

$$\mathbb{P}(X_n \in A\,;\, Y_n = 0 \,|\, X_{n-1} = x\,;\, Y_{n-1}) := Q(x, A), \tag{5.4}$$

kaikilla $n \geq 1$, $A \subset E$. Huomataan, että

$$\mathbb{P}(X_n \in A \; X_{n-1} = x, Y_{n-1} = j) = \mathbb{P}(X_n \in A \,|\, X_{n-1} = x), \quad \text{kun } j = 0, 1$$

Huomataan lisäksi, että

$$\begin{aligned} \mathbb{P}(Y_n = 1 \,|\, X_{n-1} = x\,;\, Y_{n-1}) &= \mathbb{P}(X_n \in A\,;\, Y_n = 1 \,|\, X_{n-1} = x\,;\, Y_{n-1}) \\ &\quad + \mathbb{P}(X_n \in A^c\,;\, Y_n = 1 \,|\, X_{n-1} = x\,;\, Y_{n-1}) \\ &= s(x) \left[\int_A \nu(y) dy + \int_{A^c} \nu(y) dy \right] \\ &= s(x) \int_E \nu(y) dy \\ &= s(x). \end{aligned}$$

Ehdollistamalla polulla $(X_0, Y_0), \ldots, (X_{n-1}, Y_{n-1})$ ja tapahtumalla $Y_n = 1$ huomataan, että satunnaismuuttuja X_n ei riipu enää edellisestä tilasta X_{n-1}

$$\begin{aligned} \mathbb{P}(X_n \in A \,|\, X_{n-1} = x\,;\, Y_{n-1}\,;\, Y_n = 1) &= \frac{\mathbb{P}(X_n \in A\,;\, Y_n = 1 \,|\, X_{n-1} = x\,;\, Y_{n-1})}{\mathbb{P}(Y_n = 1 \,|\, X_{n-1} = x\,;\, Y_{n-1})} \\ &= (s(x) \int_A \nu(y) dy)/(s(x)) \\ &= \int_A \nu(y) dy \\ &= (|A \cap J|)/|J|. \end{aligned}$$

Markovin unohtavaisuusominaisuuden mukaan

$$\begin{aligned}
& \mathbb{P}(X_n \in A_0, X_{n+1} \in A_1, ...; Y_{n+1} = y_1, Y_{n+2} = y_2, ... | X_0, ... X_{n-1}; \\
& Y_0, ... Y_{n-1}; Y_n = 1) \\
= \; & \mathbb{P}(X_0 \in A_0, X_1 \in A_1, \ldots ; \, Y_1 = y_1, Y_2 = y_2, \ldots \, | \, Y_0 = 1) \\
= \; & \mathbb{P}_\nu(X_0 \in A_0, X_1 \in A_1, \ldots ; \, Y_1 = y_1, Y_2 = y_2, \ldots) \qquad (5.5)
\end{aligned}$$

kaikilla $A_0, A_1, \ldots \subset E$, $y_1, y_2, \ldots \in \{0, 1\}$. Alaindeksi ν viittaa tässä alkutodennäköisyystiheysfunktioon.

Sanotaan, että Markovin ketju (X_n, Y_n) uusiutuu ajanhetkinä n, jolloin pätee $Y_n = 1$. Sanallisesti Nummelinin "*splitting*" konstruktiota voidaan kuvata nopanheitolla. Kun satunnaismuuttuja X_n vierailee joukossa I, voidaan heittää vinoutunutta noppaa, joka saa kruunan todennäköisyydellä $\beta|J|$ ja klaavan todennäköisyydellä $1 - \beta|J|$. Kruunan esiintyessä tila X_{n+1} on tasaisesti jakautunut, $X_{n+1} \sim 1/|J|$. Klaavan esiintyessä tila X_{n+1} on jakautunut siirtymäytimen Q mukaisesti.

Myöhemmin esitettävän keskeisen raja-arvolauseen todistusta varten esitetään yhtälö

$$\begin{aligned}
& \mathbb{P}(X_n \in A \mid X_0, ..., X_{n-2}; \, Y_0, ..., Y_{n-1}; \, X_{n-1} = x; \, Y_n = 0) \\
= \; & [\mathbb{P}(X_n \in A; \, Y_n = 0 \mid Y_{n-1}; \, X_{n-1} = x)] \, / \, [\mathbb{P}(Y_n = 0 \mid Y_{n-1}; \, X_{n-1} = x)] \\
= \; & Q(x, A)(1 - s(x))^{-1} \\
\leq \; & P(x, A)(1 - \beta|J|)^{-1} \\
= \; & \mathbb{P}(X_n \in A \mid X_0, ..., X_{n-1}; \, Y_0, ..., Y_{n-1}; \, X_{n-1} = x)(1 - \beta|J|)^{-1},
\end{aligned}$$

joka pätee kaikilla $n \geq 1$ ja $x \in I$.

Olkoon $1 \leq T_1 \leq T_2 \leq \ldots$ toteutuneita uusiutumisajanhetkiä:

$$\begin{aligned}
T_1 &:= \min \{n \geq 1 \, : \, Y_n = 1\}, \\
T_i &:= \min \{n > T_{i-1} \, : \, Y_n = 1\} \quad \text{kaikilla } i = 2, 3, \ldots .
\end{aligned}$$

Yhtälöstä (5.5) seuraa

$$\begin{aligned}
& \mathbb{P}(X_{T_i} \in A_0, X_{T_i+1} \in A_1, ..., X_{T_i+m-1} \in A_{m-1}; T_{i+1} - T_i = m \\
& \mid X_0, ..., X_{T_i-1}; T_1, ..., T_{i-1}; T_i = n) \\
= \; & \mathbb{P}(X_0 \in A_0, \ldots, X_{m-1} \in A_{m-1} \, ; \, T_1 = m \mid Y_0 = 1) \\
= \; & \mathbb{P}_\nu(X_0 \in A_0, \ldots, X_{m-1} \in A_{m-1} \, ; \, T_1 = m).
\end{aligned}$$

Siten satunnaisjaksot (*random blocks*)

$$\begin{aligned}\xi_0 &:= (X_0, \ldots, X_{T_1-1}\,;\, T_1)\\ \xi_i &:= (X_{T_i}, \ldots, X_{T_{i+1}-1}\,;\, T_{i+1} - T_i), \quad i = 1, 2, \ldots\end{aligned}$$

ovat riippumattomia ja samoinjakautuneita, mistä seuraa

$$\mathbb{P}(T_{i+1} - T_i = m \mid X_0, \ldots, X_{n-1};\, T_1, \ldots, T_{i-1};\, T_i = n) = \mathbb{P}_\nu(T_1 = m). \tag{5.6}$$

Kaikille funktioille $f(x)$, $x \in E$, määritellään satunnaissummat ζ_i yli ajanjakson ξ_i:

$$\zeta_0(f) := \sum_{m=0}^{T_1-1} f(X_m), \tag{5.7}$$

$$\zeta_i(f) := \sum_{m=T_i}^{T_{i+1}-1} f(X_m) \quad \text{kaikilla } i = 1, 2, \ldots. \tag{5.8}$$

Myös nämä summat ovat riippumattomia ja samoinjakautuneita.

5.4.2 Potentiaalifunktio

Uusiutumisjaksojen odotusarvon äärellisyyttä voidaan tutkia ns. potentiaalifunktiolla (*potential function*), joka määritellään tässä kuten

$$\mu(x) := \sum_{n=0}^{\infty} \nu Q^n(x).$$

Määritellään päättyvä Markovin ketju $(\widetilde{X}_n)$, jonka elinaika on $T_1 - 1$. Olkoon δ ylimääräinen tila, joka ei kuulu tilajoukkoon E. Asetetaan

$$\begin{aligned}\widetilde{X}_n &= X_n, \quad \text{kun } 0 \le n \le T - 1\\ \widetilde{X}_n &= \delta, \quad \text{kun } n \ge T.\end{aligned}$$

Markovin ketjulla $(\widetilde{X}_n)$ on alisiirtymätodennäköisyydet

$$\begin{aligned}\mathbb{P}(\widetilde{X}_n \in A \,|\, \widetilde{X}_{n-1} = x) &= \mathbb{P}(X_n \in A, T > n \,|\, X_{n-1} = x, T > n - 1)\\ &= \mathbb{P}(X_n \in A, Y_n = 0 \,|\, X_{n-1} = x, Y_{n-1} = 0)\\ &= Q(x, A).\end{aligned}$$

Olkoon tilojen X_0 ja $\widetilde{X}_0$ alkujakauma λ. Tällöin tilan $\widetilde{X}_n$ alitodennäköisyystiheysfunktio on λQ^n, koska

$$\mathbb{P}(\widetilde{X}_n \in A) = \mathbb{P}(X_n \in A, T > n) = \int_A \lambda Q^n(x) dx \tag{5.9}$$

Markovin ketjun vierailujen lukumäärän odotusarvo osajoukossa $A \subset E$ ensimmäisen satunnaisjakson ξ_0 aikana on

$$\begin{aligned} E_\nu \sum_{n=0}^{T_1-1} \delta_{X_n}(A) &= \sum_{n=0}^{\infty} \mathbb{P}_\nu(X_n \in A\,,\, T > n) \\ &\overset{\text{yhtälö (5.9)}}{=} \sum_{n=0}^{\infty} \int_A \nu Q^n(x)dx \\ &= \int_A \mu(x)dx, \end{aligned} \tag{5.10}$$

missä $\delta_{X_n}(A) = 1$, kun $X_n \in A$, ja $\delta_{X_n}(A) = 0$, kun $X_n \in A^c$. Huomataan, että uusiutumisajan odotusarvo saadaan erikoistapauksena, jossa asetetaan $A = E$, jolloin

$$M := E_\nu T = \int \mu(x)dx.$$

Yleisemmin saadaan kaikille ei-negatiivisille funktioille $f(x)$, $x \in E$, odotusarvo

$$E_\nu \zeta_0(f) = E_\nu \sum_{n=0}^{T_1-1} f(X_n) = \int \mu(x) f(x)dx. \tag{5.11}$$

5.4.3 Uusiutumisajan odotusarvon äärellisyys

Tässä kappaleessa näytetään, että uusiutumisajan odotusarvo on äärellinen.

Kuvatkoon tapahtuma $\{T \leq n\}$, $n \geq 1$, aikaa, joka on kulunut edellisen uusiutumisen jälkeen ennen ajanhetkeä n. Olkoon siis

$$L_n := \min\{0 \leq k \leq n-1 \,:\, Y_{n-k} = 1\}\,,$$

viimeisen regeneraation aika ennen ajanhetkeä n. Tällöin

$$\{T \leq n\} = \bigcup_{k=0}^{n-1} \{L_n = k\}\,.$$

Tällöin saadaan identiteetti

$$\begin{aligned}
\mathbb{P}_\lambda(X_n \in A) &= \mathbb{P}_\lambda(X_n \in A, T > n) + \mathbb{P}_\lambda(X_n \in A, T \leq n) \\
&= \mathbb{P}_\lambda(X_n \in A, T > n) + \sum_{k=0}^{n-1} \mathbb{P}_\lambda(L_n = k, X_n \in A) \\
&= \mathbb{P}_\lambda(X_n \in A, T > n) \\
&\quad + \sum_{k=0}^{n-1} \mathbb{P}_\lambda(Y_{n-k} = 1, Y_{n-k+1} = 0, \ldots, Y_k = 0;\, X_n \in A) \\
&= \mathbb{P}_\lambda(X_n \in A, T > n) \\
&\quad + \sum_{k=0}^{n-1} \mathbb{P}_\lambda(Y_{n-k} = 1)\mathbb{P}_\nu(Y_{n-k+1} = 0, \ldots, Y_k = 0;\, X_n \in A) \\
&= \mathbb{P}_\lambda(X_n \in A, T > n) + \sum_{k=0}^{n-1} \mathbb{P}_\lambda(Y_{n-k} = 1)\mathbb{P}_\nu(T > k;\, X_k \in A) \\
&= \mathbb{P}_\lambda(X_n \in A, T > n) + \sum_{k=0}^{n-1} \int \lambda \mathbb{P}^{n-k-1}(y) s(y) dy \int_A \nu Q^k(x) dx,
\end{aligned}$$

joka on voimassa kaikilla alkutodennäköisyystiheysfunktioilla λ ja kaikilla $n \geq 1$, $A \subset E$.

Jos alkutodennäköisyystiheysfunktio on tasapainojakauma, eli $\lambda = \pi$, niin identiteetti on muotoa

$$\begin{aligned}
\pi(A) &= \int_A \pi(y) dy \\
&= \mathbb{P}_\pi(X_n \in A,\, T > n) + \int \pi(x) s(x) dx \sum_{k=0}^{n-1} \int_A \nu Q^k(y) dy. \qquad (5.12)
\end{aligned}$$

Asettamalla $A = E$ saadaan

$$1 = \int \pi(y) dy = \mathbb{P}_\pi(T > n) + \int \pi(x) s(x) dx \sum_{k=0}^{n-1} \int \nu Q^k(y) dy.$$

Antamalla n:n kasvaa kohti ääretöntä edellisessä yhtälössä saadaan

$$\begin{aligned}
1 &= \mathbb{P}_\pi(T = \infty) + \int \pi(x) s(x) dx \sum_{k=0}^{\infty} \int \nu Q^k(y) dy \\
&= \mathbb{P}_\pi(T = \infty) + \int \pi(x) s(x) dx \int \sum_{k=0}^{\infty} \nu Q^k(y) dy \\
&= \mathbb{P}_\pi(T = \infty) + M \int \pi(x) s(x) dx. \qquad (5.13)
\end{aligned}$$

Hypoteeseista (5.8) ja (5.9) seuraa

$$\int \pi(x) s(x) dx = \beta \mid J \mid \int \pi(x) \sum_{n=1}^{\infty} 2^{-n} P^n(x, I) dx > 0, \qquad (5.14)$$

jolloin yhdessä yhtälön (5.13) voidaan päätellä uusiutumisajan odotusarvon M äärellisyys:

$$M := E_\nu T = \int \mu(x) dx < \infty. \qquad (5.15)$$

Siten pätee myös

$$\mathbb{P}_\nu(T < \infty) = \int \mu(x)s(x)dx = 1. \tag{5.16}$$

Yhtälöstä (5.15) seuraa, että funktio $M^{-1}\mu(x)$ on todennäköisyystiheysfunktio.

5.4.4 Invariantti jakauma

Tässä kappaleessa todistetaan, että todennäköisyystiheysfunktio $M^{-1}\mu(x)$ on invariantti ja yksikäsitteisesti määrätty.

Lause 5.11. *Todennäköisyystiheysfunktio $M^{-1}\mu(x)$ on invariantti*

Todistus.

$$\begin{aligned}
M^{-1}\mu \quad &= \quad M^{-1}\sum_{n=0}^{\infty}\nu Q^n(y) \\
&= \quad M^{-1}\left[\nu(y) + \sum_{n=1}^{\infty}\nu Q^n(y)\right] \\
&= \quad M^{-1}\left[\nu(y)\cdot 1 + \sum_{n=0}^{\infty}(\nu Q^n)Q(y)\right] \\
&\overset{\text{yhtälö (5.16)}}{=} \quad M^{-1}\left[\nu(y)\int \mu(x)s(x)dx + \mu Q(y)\right] \\
&\overset{\text{yhtälö (5.2)}}{=} \quad M^{-1}\mu P(y).
\end{aligned}$$

□

Lause 5.12. *Invariantti todennäköisyystiheysfunktio $M^{-1}\mu(x)$ on yksikäsitteinen, eli*

$$\pi = M^{-1}\mu. \tag{5.17}$$

Todistus. Annetaan $n \to \infty$ yhtälössä (5.12), jolloin saadaan epäyhtälö

$$\int_A \pi(y)dy \quad \geq \quad \int \pi(x)s(x)dx \int_A \mu(y)dy,$$

joka pätee kaikilla $A \subset E$. Siirtämällällä kaikki termit samalle puolelle ja muokkaamalla integraalilauseketta saadaan

$$\int_A \left[\pi(y) - \mu(y)\int \pi(x)s(x)dx\right]dy \geq 0, \tag{5.18}$$

Määritellään $\lambda(y)$ kuten

$$\lambda(y) := \pi(y) - \mu(y)\int \pi(x)s(x)dx, \tag{5.19}$$

jolloin yhtälö (5.18) saa muodon

$$\int_A \lambda(y)dy \geq 0.$$

Tällöin pätee myös $\lambda(y) > 0$ melkein kaikkialla.

Yhtälön (5.17) avulla saadaan

$$\begin{aligned} \int \pi(x)s(x)dx &= \int M^{-1}\mu(x)s(x)dx \\ &= M^{-1}. \end{aligned} \tag{5.20}$$

Jos $\lambda(y) \equiv 0$, niin edellisestä yhtälöstä sekä yhtälöstä (5.19) seuraa

$$\pi(y) \equiv \mu(y) \int \pi(x)s(x)dx \equiv M^{-1}\mu(y) \quad \text{melkein kaikkialla.}$$

Jos taas pätee $\int \lambda(y)dy > 0$, niin funktio

$$\frac{\lambda(x)}{\int \lambda(y)dy}$$

on invariantti todennäköisyystiheysfunktio, jolle yhtälön (5.16) nojalla pätee

$$\begin{aligned} & & \int \mu(x)s(x)dx &= 1 \\ & \overset{\text{yhtälö (5.19)}}{\Leftrightarrow} \frac{1}{\int \pi(y)s(y)dy} & \int \pi(x)s(x) - \lambda(x)s(x)dx &= 1 \\ & & \Leftrightarrow \frac{\int \lambda(x)s(x)dx}{\int \pi(y)s(y)dy} &= 0 \\ & & \Leftrightarrow \int \lambda(x)s(x)dx &= 0. \end{aligned}$$

Tämä on kuitenkin ristiriidassa yhtälön (5.14) kanssa. Siis $\pi = M^{-1}\mu$ on yksikäsitteinen.

□

5.4.5 Rekurssiivisuus

Lause 5.13. *Markovin ketju on rekurssiivinen, eli*

$$\mathbb{P}(T_i < \infty \mid X_0 = x) = 1 \quad \textit{kaikilla } x \in S \textit{ ja } i = 1, 2, \ldots.$$

Uusiutuminen tapahtuu siis äärettömän usein todennäköisyydellä 1.

Todistus. Induktiolla ja yhtälöiden (5.6) ja (5.16) avulla voidaan todistaa, että yhtälö

$$\mathbb{P}(T_i < \infty \mid X_0 = x) = \mathbb{P}(T_1 < \infty \mid X_0 = x)$$

pätee mielivaltaisella alkutilalla $x \in E$ ja $i \geq 1$. Siten riittää todistaa, että lause pätee uusiutumisajalla T_1.

Yhtälöistä (5.13) ja (5.20) seuraa

$$\begin{aligned} 1 &= \mathbb{P}_\pi(T_1 = \infty) + M^{-1}M \\ \Leftrightarrow 1 &= 1 - \mathbb{P}_\pi(T_1 < \infty) + 1 \\ \Leftrightarrow 1 &= \int \mathbb{P}(T_1 < \infty \mid X_0 = x)\pi(x)dx. \end{aligned}$$

Tällöin melkein kaikilla $x \in S$ pätee

$$\mathbb{P}(T_1 < \infty \mid X_0 = x) = 1. \tag{5.21}$$

Määritellään kaikilla alkutiloilla $X_0 = x \in S$ funktio $h_\infty(x)$ kuten

$$h_\infty(x) := \mathbb{P}(T_1 = \infty \mid X_0 = x).$$

Se voidaan kuvata myös raja-arvon avulla seuraavasti

$$h_\infty(x) = \lim_{n\to\infty} \mathbb{P}(T_1 > n \mid X_0 = x) = \lim_{n\to\infty} Q^n(x, E).$$

Funktio $h_\infty(x)$ voidaan ilmaista myös siirtymäytimen $Q(x, y)$ ja funktion $h_\infty(y)$ avulla:

$$\int Q(x, dy)h_\infty(y) \equiv h_\infty(x)dy.$$

Yhtälön (5.16) nojalla

$$\int \nu(y)h_\infty(y)dy = 0,$$

jolloin siirtymäytimen Q määritelmän (5.2) mukaan

$$\begin{aligned} h_\infty(x) &\equiv \int Q(x, dy)h_\infty(y) \\ &\equiv \int P(x, dy)h_\infty(y) - s(x)\int \nu(y)h_\infty(y)dy \\ &\equiv \int P(x, dy)h_\infty(y). \end{aligned}$$

Edelleen yhtälö voidaan ilmaista todennäköisyyden $r(x)$ avulla

$$\int p(x, y)h_\infty(y)dy \equiv (1 - r(x))h_\infty(x).$$

Oletetaan, että eräällä $x_0 \in S$ $h_\infty(x_0)$ on aidosti positiivinen. Kappaleessa (5.3) todettiin, että $r(x) < 1$ kaikilla $x \in E$, jolloin saadaan

$$\int p(x_0, y)h_\infty(y)dy > 0$$

Hypoteesin (5.9) kohdan (ii) mukaan

$$\int_S p(x_0, y)h_\infty(y)dy = \int p(x_0, y)h_\infty(y)dy > 0,$$

josta seuraa

$$\int_S h_\infty(y)dy > 0.$$

Tämä on kuitenkin ristiriidassa yhtälön (5.21) kanssa, joten lause pitää paikkansa. □

5.4.6 SLL:n todistus

Olkoon satunnaissummat ζ_i, $i = 1, 2, \ldots$ määritelty kuten yhtälöissä (5.7) ja (5.8). Koska satunnaismuuttujat $\zeta_0, \varsigma_1, \ldots$ ovat riippumattomia ja satunnaismuuttujat $\zeta_1, \zeta_2, \ldots$ ovat samoinjakautuneita, seuraa tavallisesta suurten lukujen laista ja satunnaismuuttujan ζ_0 äärellisyydestä (lause 5.13)

$$\begin{aligned}\lim_{i\to\infty} i^{-1}\sum_{j=0}^{i}\zeta_j &= E\zeta_1 = E_\nu\zeta_0 \overset{\text{yhtälö (5.11)}}{=} \int f(x)\mu(x)dx. \\ &= M\int f(x)\pi(x)dx \\ &= M\pi(f) \quad \text{todennäköisyydellä 1.} \end{aligned} \tag{5.22}$$

Vastaavasti yhtälöstä (5.15) seuraa

$$\begin{aligned}\lim_{i\to\infty} i^{-1}T_i &= E(T_2 - T_1) = E_\nu T_1 \\ &= M \quad \text{todennäköisyydellä 1.}\end{aligned} \tag{5.23}$$

Olkoon $N(n)$, $n = 1, 2, \ldots$, niiden uusiutumiskertojen T_i lukumäärä ennen ajanhetkeä n. Lauseesta (5.13) seuraa, että $N(n) \to \infty$, kun $n \to \infty$. Tällöin täytyy

$$\begin{aligned}\lim_{n\to\infty} n^{-1}N(n) &= \lim_{n\to\infty}(T_{N(n)})^{-1}N(n) \\ &\overset{\text{yhtälö (5.23)}}{=} M^{-1} \quad \text{todennäköisyydellä 1.}\end{aligned} \tag{5.24}$$

Kirjoitetaan summa $S_n := \sum_{m=0}^{n-1} f(X_m)$ muodossa

$$S_n = \sum_{j=0}^{N(n)-1}\zeta_j + \zeta'_{N(n)},$$

missä

$$\begin{aligned}\zeta'_{N(n)} &:= \sum_{m=T_{N(n)}}^{n-1} f(X_m), \quad \text{jos } T_{N(n)} \le n-1 \\ \zeta'_{N(n)} &:= 0 \quad \text{jos } T_{N(n)} = n.\end{aligned}$$

Nähdään, että

$$\mid \zeta'_{N(n)} \mid \le \sum_{m=T_{N(n)}}^{T_{N(n)+1}-1} \mid f(X_m) \mid,$$

missä satunnaismuuttuja $\sum_{m=T_{N(n)}}^{T_{N(n)+1}-1} \mid f(X_m) \mid$ on samoinjakautunut kuin äärellinen satunnaismuuttuja $\sum_{m=0}^{T-1} \mid f(X_m) \mid$. Tällöin voidaan päätellä, että

$$\lim_{n\to\infty} n^{-1}\zeta'_{N(n)} = 0 \quad \text{todennäköisyydellä 1.} \tag{5.25}$$

Lopulta saadaan

$$\begin{aligned}
\lim_{n\to\infty} n^{-1}S_n &\overset{\text{yhtälö (5.25)}}{=} \lim_{n\to\infty} n^{-1}\sum_{j=0}^{N(n)-1}\zeta_j \\
&= \lim_{n\to\infty} n^{-1}(N(n)-1)\frac{\sum_{j=0}^{N(n)-1}\zeta_j}{N(n)-1} \\
&\overset{\text{yhtälöt (5.22) ja (5.24)}}{=} M^{-1}M\pi(f) - M\pi(f)\lim_{n\to\infty} n^{-1} \\
&= \pi(f),
\end{aligned}$$

mikä on sama kuin teoreeman (5.10) väite.

5.5 Keskeinen raja-arvolause Markovin ketjulle

Määritelmä 5.14. Epätyhjä joukko $K \subset E$ on heikosti pieni (*weakly small*), jos se voidaan peittää äärellisellä määrällä pieniä joukkoja $\{I_k\}$.

Hypoteesi 5.15. On olemassa heikosti pieni joukko $K \subset E$, äärellinen ja ei-negatiivinen π-integroituva funktio $h(x) \geq 0$ sekä vakiot $C < \infty$ ja $\delta > 0$ siten, että

$$E(h(X_1) \mid X_0 = x) \leq C \quad \text{kaikilla } x \in K, \text{ ja} \tag{5.26}$$

$$E(h(X_1) - h(X_0) \mid X_0 = x) \leq -\delta \quad \text{kaikilla } x \in K^c. \tag{5.27}$$

Teoreema 5.16. *Oletetaan, että hypoteesit (5.8), (5.9) ja (5.15) ovat voimassa. Tällöin*

$$n^{\frac{1}{2}}(\widehat{\pi}_n(f) - \pi(f)) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2)$$

kaikilla rajoitetuilla funktioilla f ja alkutiloilla $X_0 = x \in S$.

5.5.1 Ergodisuus

Määritelmä 5.17. Markovin ketju (X_n) on ergodinen toisen asteen suhteen, jos ensimmäisen uusiutumishetken toinen momentti ehdolla alkutodennäköisyystiehysfunktio ν on äärellinen, eli

$$E_\nu T_1^2 < \infty.$$

Lause 5.18. *Hypoteesista (5.15) seuraa Markovin ketjun (X_n) ergodisuus toisen asteen suhteen.*

Lauseen todistusta varten todistetaan ensin kaksi lemmaa. Niitä varten on tarpeen tehdä muutamia merkintöjä.

Olkoon $\tau \geq 1$ ensimmäinen ajankohta, jolloin Markovin ketju (X_n) saa arvon joukossa K:

$$\tau := \min\{n \geq 1 : X_n \in K\}.$$

Yleisemmin voidaan määritellä $0 \leq \tau_0 \leq \tau_1 \leq \ldots$ olemaan ajankohtia, jolloin Markovin ketju (X_n) saa arvon joukossa K:

$$\begin{aligned} \tau_0 &:= \min\{n \geq 0 : X_n \in K\}, \\ \tau_i &:= \min\{n > \tau_{n-1} : X_n \in K\}, \quad \text{kaikilla } i = 1, 2, \ldots. \end{aligned}$$

Lemma 5.19. *Olkoon hypoteesit (5.9) ja (5.15) voimassa. Tällöin*

$$\begin{aligned} M := \max_{x \in K} E(\tau \mid X_0 = x) &< \infty \quad ja \\ E_\pi \tau_0 &< \infty. \end{aligned}$$

Todistus. Määritellään päättyvä Markovin ketju $(X_{K;n})$, jolle pätee

$$\begin{aligned} X_{K;n} &= X_n \quad \text{kaikilla } 0 \leq n \leq \tau_0, \\ X_{K;n} &= \delta \quad \text{kaikilla } n \geq \tau_0 + 1, \end{aligned}$$

missä δ tarkoittaa joukkoon E kuulumatonta tilaa. Tämän Markovin ketjun siirtymätodennäköisyydet ovat

$$\begin{aligned} P_K(x, A) &:= P(x, A) \quad \text{kaikilla } x \in K^c, A \subset E, \\ P_K(x, A) &:= 0 \quad \text{kaikilla } x \in K, A \subset E. \end{aligned}$$

Epäyhtälö (5.27) voidaan ilmaista myös muodossa

$$\begin{aligned} E(h(X_1) - h(X_0) \mid X_0 = x) &\leq -\delta \\ \int [h(y) - h(x)] P_K(x, y) dy &\leq -\delta \\ \int h(y) P_K(x, y) dy - h(x) \int P_K(x, y) dy &\leq -\delta \\ \delta P_K(x, E) + \int h(y) P_K(x, y) dy &\leq h(x). \end{aligned}$$

Induktiolla on helppo osoittaa, että edellinen yhtälö pätee myös useammalle siirtymätodennäköisyyden potenssille:

$$\delta \sum_{k=1}^{n} P_K^k(x, E) + \int h(y) P_K^n(x, y) dy \leq h(x) \quad \text{kaikilla } n = 1, 2, \ldots.$$

Antamalla $n \to \infty$ edellisen yhtälön integraalilauseke lähestyy nollaa, jolloin saadaan

$$\begin{aligned} h(x) &\geq \delta \sum_{k=1}^{\infty} P_K^k(x, E) \\ &= \delta \sum_{k=0}^{\infty} \mathbb{P}(\tau_0 \geq k \mid X_0 = x) \\ &= \delta(1 + E(\tau_0 \mid X_0 = x)). \end{aligned} \tag{5.28}$$

Markov ominaisuuden avulla saadaan

$$\begin{aligned} E(\tau \mid X_0 = x) &= \int P(x, dy)(1 + E(\tau_0 \mid X_0 = y)) \\ &\overset{\text{yhtälö (5.28)}}{\leq} \delta^{-1} \int P(x, y) h(y) dy \\ &= \delta^{-1} E(h(X_1) \mid X_0 = x) \\ &\overset{\text{yhtälö (5.26)}}{\leq} \delta^{-1} C < \infty \quad \text{kaikilla } x \in K. \end{aligned}$$

Tällöin täytyy myös

$$\max_{x \in K} E(\tau \mid X_0 = x) < \infty.$$

Integroimalla yli yhtälön (5.28) saadaan

$$\begin{aligned} \int \pi(x)(1 + E(\tau_0 \mid X_0 = x)) dx &\leq \delta^{-1} \int \pi(x) h(x) dx \\ E_\pi \tau_0 &\leq \delta^{-1} \pi(h) < \infty. \end{aligned}$$

Täten lemman molemmat väitteet tulevat todistetuksi. □

Lemma 5.20. *Olkoon $K = I$ pieni joukko, joka toteuttaa lemman (5.19) väitteet. Tällöin Markovin ketju on ergodinen toisen asteen suhteen.*

Todistus. Riittää osoittaa, että odotusarvo $E_\pi T$ on äärellinen. Tarkemmat perustelut on luettavissa Esa Nummelinin kirjoituksesta *MC's for MCMC'ists* lehdestä *International Statistical Review (2002).*

Oletetaan, että uusiutuminen eli tapahtuma $Y_n = 1$ voi tapahtua vain sellaisina ajanhetkinä n, joille pätee $X_{n-1} \in I$. Tällöin ensimmäinen uusiutumisajankohta on muotoa $T_1 = \tau_\kappa + 1$, missä

$$\kappa := \min \left\{ k \geq 0 \,:\, Y_{\tau_{k+1}} = 1 \right\}.$$

Siirtymätodennäköisyyden määritelmästä (5.3) seuraa

$$\mathbb{P}(T = n \mid X_0, \ldots, X_{n-2};\, X_{n-1} = x, T > n-1) \equiv \beta > 0, \quad \text{kun } x \in I,$$

ja riippumatta ajanhetkestä $n \geq 1$. Tästä seuraa yhtälö

$$\mathbb{P}(\kappa = k+1 \mid \kappa > k) \equiv \mathbb{P}(T = n \mid X_{n-1} = x, \tau_{k+1} = n-1, T > n-1) \equiv \beta,$$

riippumatta muuttujasta $x \in I$ ja ajanhetkestä $n \geq 1$. Siten κ on jakautunut kuten geometrinen jakauma parametrina β. Nyt saadaan

$$\begin{aligned} & E(\tau_{k+1} - \tau_k \mid X_0, ..., X_{n-2};\, X_{n-1} = x;\, \tau_k = n-1;\, Y_0, ..., Y_{n-1};\, Y_n = 0) \\ \leq\ & (1 - \beta|J|)^{-1} E(\tau_{k+1} - \tau_k \mid X_{n-1} = x;\, \tau_k = n-1) \\ =\ & (1 - \beta|J|)^{-1} E(\tau \mid X_0 = x) \\ \leq\ & (1 - \beta|J|)^{-1} M. \end{aligned} \tag{5.29}$$

Tapahtumasta $\kappa > k$ seuraa se, että jollakin $n \geq 1$ pätee $X_{n-1} = x$, missä $x \in I$, $\tau_k = n-1$ ja $Y_n = 0$. Tästä ja yhtälöstä (5.29) seuraa

$$\begin{aligned} E(\tau_{k+1} - \tau_k \mid \kappa > k) &= E(\tau_{k+1} - \tau_k \mid X_{n-1} = x;\, \tau_k = n-1;\, Y_n = 0) \\ &\leq (1 - \beta|J|)^{-1} M. \end{aligned}$$

Ensimmäinen uusiutumisaika T_1 voidaan kirjoittaa kuten

$$\begin{aligned} T_1 &= 1 + \sum_{k=0}^{\infty} \tau_k 1_{\{\kappa = k\}} \\ &= 1 + \tau_0 + \sum_{k=0}^{\infty} (\tau_{k+1} - \tau_k) 1_{\{\kappa > k\}}. \end{aligned}$$

Olkoon $\lambda = \pi$, jolloin saadaan lemman (5.19) nojalla

$$\begin{aligned} E_\pi T_1 &= 1 + E_\pi \tau_0 + \sum_{k=0}^{\infty} E_\pi(\tau_{k+1} - \tau_k \mid \kappa > k) P(\kappa > k) \\ &\leq 1 + E_\pi \tau_0 + (1 - \beta|J|)^{-1} M \sum_{k=0}^{\infty} P(\kappa > k) \\ &= 1 + E_\pi \tau_0 + (1 - \beta|J|)^{-1} M E\kappa < \infty. \end{aligned}$$

□

Nyt voidaan todistaa Markovin ketjun ergodisuus.

Todistus. Joukon K oletettiin olevan heikosti pieni. Tällöin on olemassa äärellinen määrä pieniä joukkoja I_k, jotka peittävät joukon K ja joille pätee

$$p(x \to y) \geq \beta, \quad x \in I_k, y \in J_k$$

kaikilla $k = 1, \ldots, d$. Kun Markovin ketju vierailee joukossa K, voidaan heittää vinoutunutta noppaa, joka tuottaa kruunan todennäköisyydellä β ja klaavan todennäköisyydellä $1 - \beta$. Jos tulos on kruuna, niin riippuen siitä, mihin joukkoon I_k nykyinen tila kuuluu, seuraava tila on jakautunut kuten

$$\nu_k := \text{tasainen todennäköisyystiheysfunktio joukolle } J_k.$$

Olkoon τ_K ajankohta, jolloin ensimmäisen kerran Markovin ketju vierailee joukossa K. Lemman (5.19) nojalla odotusarvot $E_\pi \tau_K$ ja $\max_{x \in K} E(\tau_K \mid X_0 = x)$ ovat äärellisiä. Olkoon τ_{I_k} ajankohta, jolloin Markovin ketju vierailee ensimmäisen kerran joukossa I_k. Tällöin käyttämällä uudestaan geometrisen jakauman argumenttia löydetään indeksi k siten, että odotusarvot $E_\pi \tau_{I_k}$ ja $\sup_{x \in I_k} E(\tau_{I_k} \mid X_0 = x)$ ovat äärellisiä. Näin löytyi pieni joukko I_k, jolle pätee lemman (5.19) väitteet. Siten lemman (5.20) nojalla Markovin ketju on ergodinen toisen asteen suhteen. □

Lause 5.21. *(Egorovin lause). Olkoon (f_n) jono reaaliarvoisia funktioita mitta-avaruudessa $(\Omega, \mathcal{F}, \mu)$, jotka suppenevat kohti funktiota f μ-melkein varmasti. Tällöin kaikilla $\varepsilon > 0$ on olemassa mitallinen joukko $B \subset \Omega$ siten, että $\mu(B) < \varepsilon$. Lisäksi funtiojono (f_n) suppenee tasaisesti kohti funktiota f joukossa $\Omega \backslash B$.*

Todistus. Olkoon n ja k luonnollisia lukuja. Määritellään

$$E_{n,k} \bigcup_{m \geq n} \left\{ x \in \Omega \mid |f_m(\omega) - f(\omega)| \geq \frac{1}{k} \right\}.$$

Kun n kasvaa, joukot $E_{n,k}$ pienenevät. Siten oletus funktiojonon (f_n) suppenemisesta kohti funktiota f μ-melkein varmasti tuottaa

$$\mu(\bigcap_{n \in \mathbb{N}} E_{n,k}) = 0 \quad \text{kaikilla } k.$$

Tästä seuraa myös mitan μ jatkuvuus. Tällöin on olemassa luonnollinen luku n_k siten, että

$$\mu(E_{n_k,k}) < \frac{\varepsilon}{2^k}.$$

Määritellään

$$B = \bigcup_{k \in \mathbb{N}} E_{n_k,k},$$

joka on niiden perusjoukon Ω alkioiden ω joukko, joilla $f_m(\omega)$, $m > n_k$, lähestyy liian hitaasti kohti funktiota f ainakin yhdellä k:n arvolla. Siten joukolle $\Omega \backslash B$ pätee tasainen suppeneminen.

Additiivisuudesta sekä geometrisen sarjan suppenemisesta seuraa

$$\mu(B) \leq \sum_{k \in \mathbb{N}} \mu(E_{n_k,k}) < \sum_{k \in \mathbb{N}} \frac{\varepsilon}{2^k} = \varepsilon.$$

[13, s.49-50, 2007.] □

Lause 5.22. *(Kolmogorovin epäyhtälö.) Olkoon $X_1, X_2, \ldots, X_n$ riippumattomia satunnaismuuttujia siten, että $E(X_j) = 0$ ja $E(X_j^2) < \infty$, kun $j = 1, 2, \ldots, n$. Tällöin pätee*

$$P\left\{ \max_{1 \leq k \leq n} \left| \sum_{j=1}^{k} X_j \right| \geq x \right\} \leq \frac{\sum_{j=1}^{n} E(X_j^2)}{x^2}, \quad x > 0.$$

Todistus. Merkitään $S_k = \sum_{j=1}^{k} X_j$. Määritellään tapahtumat

$$A_k = \{|S_k| \geq x, |S_j| < x, j = 1, 2, \ldots, k-1\},$$

jotka ovat kaikilla $k = 1, 2, \ldots, n$ pistevieraita. Tällöin funktioilla 1_{A_k} pätee

$$0 \leq \sum_{k=1}^{n} 1_{A_k} \leq 1.$$

Koska satunnaismuuttujat X_k, $k = 1, 2, \ldots, n$, ovat riippumattomia ja $E(X_k) = 0$, niin

$$E(S_n^2) = \sum_{k=1}^{n} E(X_k^2) + 2\sum_{k=1}^{n}\sum_{m=1}^{n-k} \underbrace{E(X_k X_{m+k})}_{=0} = \sum_{k=1}^{n} E(X_k^2).$$

Tällöin saadaan

$$\sum_{k=1}^{n} E(X_k^2) = E(S_n^2) \geq E(S_n^2 \sum_{k=1}^{n} 1_{A_k}) = \sum_{k=1}^{n} E(S_n^2 1_{A_k}).$$

Satunnaissummat $S_n - S_k$, missä $k < n$, ja $S_k 1_{A_k}$ ovat riippumattomia, joten

$$E((S_n - S_k)S_k 1_{A_k}) = E(S_n - S_k)E(S_k 1_{A_k}) = \sum_{j=k+1}^{n} E(X_j) = 0.$$

Koska kaikilla $\omega \in A_k$ pätee $S_k^2(\omega) \geq x^2$, niin

$$\begin{aligned} E(S_n^2 1_{A_k}) &= E((S_n - S_k + S_k)^2 1_{A_k}) \\ &= E((S_n - S_k)^2 1_{A_k}) + 2E((S_N - S_k)S_k 1_{A_k}) + E(S_k^2 1_{A_k}) \\ &\geq 0 + 0 + x^2 P(A_k). \end{aligned}$$

Koska

$$\sum_{k=1}^{n} \mathbb{P}(A_k) = \mathbb{P}(\bigcup_{k=1}^{n} A_k) = \mathbb{P}(\max_{1 \leq k \leq n} |S_k| \geq x),$$

niin väite saadaan seuraavasti

$$\sum_{k=1}^{n} E(X_k^2) \geq \sum_{k=1}^{n} E(S_n^2 1_{A_k}) \geq x^2 \sum_{k=1}^{n} \mathbb{P}(A_k) = x^2 \mathbb{P}(\max_{1 \leq k \leq n} |S_k| \geq x).$$

□

5.5.2 KRL:n todistus

Lause todistetaan tässä vain erikoistapaukselle, jossa rajoitetulle funktiolle $f(x)$ pätee $\pi(f) = 0$. Tällöin keskeisen raja-arvolauseen väite on muotoa

$$n^{\frac{1}{2}} \widehat{\pi}_n(f) \xrightarrow{d} \mathcal{N}(0, \sigma_f^2).$$

Merkitään jälleen satunnaissummia $\zeta_i(f)$ lyhyesti ζ_i. Olkoon myös $M := E_\nu T_1$. Aiemmin on osoitettu, että $n^{-1}\varsigma_0 \to 0$, $n^{-1}\zeta'_{N(n)} \to 0$ ja $n^{-1}N(n) \to M$ todennäköisyydellä 1. Siten riittää osoittaa, että

$$n^{-\frac{1}{2}} \sum_{j=1}^{N(n)-1} \zeta_j \xrightarrow{d} \mathcal{N}(0, M\sigma_f^2).$$

Olkoon $\varepsilon, \eta > 0$ mielivaltaisia vakioita. Egorovin lauseen nojalla on olemassa tapahtuma Λ, jonka komplementille pätee $P(\Lambda^c) < \eta$. Lisäksi funktiojono $(n^{-1}N(n))$ suppenee tasaisesti kohti lukua M^{-1} joukossa Λ, eli on olemassa indeksi n_η siten, että ehdolla tapahtuma Λ

$$|n^{-1}N(n) - M^{-1}| < \eta \quad \text{kaikilla } n > n_\eta.$$

Edellistä epäyhtälöä muokkaamalla saadaan

$$n(M^{-1} - \eta) < N(n) < n(M^{-1} + \eta).$$

kaikilla $n > n_\eta$ ja ehdolla tapahtuma Λ. Tällöin voidaan approksimoida kuten

$$\left| \sum_{j=1}^{N(n)-1} \zeta_j - \sum_{j=1}^{[n(M^{-1}-\eta)]} \zeta_j \right| \leq \max_{n(M^{-1}-\eta)<j<n(M^{-1}+\eta)} \left| \sum_{n(M^{-1}-\eta)<i\leq j} \zeta_i \right|,$$

missä merkintä $[n(M^{-1} - \eta)]$ tarkoittaa luvun $n(M^{-1} - \eta)$ kokonaisosaa. Kun käytetään tavallista keskeistä raja-arvolausetta riippumattomille satunnaismuuttujille, saadaan

$$n^{-\frac{1}{2}} \sum_{j=1}^{[n(M^{-1}-\eta)]} \zeta_j \xrightarrow{d} \mathcal{N}(0, E_\nu \zeta_0^2),$$

missä varianssi on saatu seuraavasti:

$$Var_\nu \zeta_0^2 = E_\nu \zeta_0^2 - (E_\nu \zeta_0)^2 \overset{\text{yhtälö (5.22)}}{=} E_\nu \zeta_0^2 - (M\pi(f))^2 = E_\nu \zeta_0^2.$$

Huomataan, että toisen asteen ergodisuuden nojalla varianssi $E_\nu \zeta_0^2$ on äärellinen:

$$E_\nu \zeta_0^2 \leq E_\nu (\overbrace{\sup_{x\in E} |f(x)| + \ldots + \sup_{x\in E} |f(x)|}^{T_1 \text{ kpl}})^2 = E_\nu T_1^2 \sup_{x\in E} |f(x)|^2 < \infty.$$

Kolmogorovin epäyhtälöstä saadaan

$$\begin{aligned} \mathbb{P}\left(\max_{n(a-\eta)<j<n(a+\eta)} \left| \sum_{n(a-\eta)<i\leq j} \zeta_i \right| > n^{\frac{1}{2}}\varepsilon \right) &\leq \mathbb{P}\left(\max_{1\leq j\leq n(a+\eta)} \left| \sum_{i=1}^{j} \zeta_i \right| > n^{\frac{1}{2}}\varepsilon \right) \\ &\leq n^{-1}\varepsilon^{-2} E_\nu \zeta_0^2 \\ &\leq 2n\eta \cdot n^{-1}\varepsilon^{-2} E_\nu \zeta_0^2 \\ &= 2\eta\varepsilon^{-2} E_\nu \zeta_0^2. \end{aligned} \tag{5.30}$$

Aina pätee

$$\mathbb{P}(A) \leq \mathbb{P}(B^c) + \mathbb{P}(A|B).$$

kaikilla joukoilla A ja B[3]. Tämän ja yhtälön (5.30) avulla pätee

$$\begin{aligned}
& \mathbb{P}\left(n^{-\frac{1}{2}} \left|\sum_{j=1}^{N(n)} \zeta_j - \sum_{j=1}^{[n(a-\eta)]} \zeta_j\right| > \varepsilon\right) \\
\leq \quad & \mathbb{P}(\Lambda^c) + \mathbb{P}\left(n^{-\frac{1}{2}} \left|\sum_{j=1}^{N(n)} \zeta_j - \sum_{j=1}^{[n(a-\eta)]} \zeta_j\right| > \varepsilon \mid \Lambda\right) \\
\leq \quad & \mathbb{P}(\Lambda^c) + \mathbb{P}(\max_{n(a-\eta)<j<n(a+\eta)} \left|\sum_{n(a-\eta)<i\leq j} \zeta_i\right| > n^{\frac{1}{2}}\varepsilon) \\
\leq \quad & \eta + 2\eta\varepsilon^{-2} E_\nu \zeta_0^2,
\end{aligned}$$

joka saadaan pienemmäksi kuin mikä tahansa annettu luku $\varepsilon' > 0$ valitsemalla tarpeeksi pieni η:n arvo. Tällöin siis saadaan

$$n^{-\frac{1}{2}} \left| \sum_{j=1}^{N(n)} \zeta_j - \sum_{j=1}^{[n(a-\eta)]} \zeta_j \right| \xrightarrow{P} 0,$$

jolloin $n^{-\frac{1}{2}} \sum_{j=1}^{N(n)} \zeta_j$ suppenee jaukaumaltaan kohti normaalijakaumaa:

$$\lim_{n\to\infty} n^{-\frac{1}{2}} \sum_{j=1}^{N(n)} \zeta_j = \lim_{n\to\infty} n^{-\frac{1}{2}} \sum_{j=1}^{[n(a-\eta)]} \zeta_j = \mathcal{N}(0, E_\nu \zeta_0^2).$$

Varianssi $E_\nu \zeta_0^2$ saadaan johdettua seuraavasti

$$\begin{aligned}
E_\nu \zeta_0^2 &= E_\nu (\sum_{m=0}^{T_1-1} f(X_m))^2 \\
&= E_\nu \sum_{m=0}^{T_1-1} (f(X_m))^2 + 2E_\nu \sum_{k=0}^{T_1-1} f(X_k) \sum_{m=1}^{T_1-1-k} f(X_{k+m}) \\
&= E_\nu \sum_{m=0}^{T_1-1} (f(X_m))^2 \\
&\quad +2 \int \sum_{k=0}^{\infty} E_\nu \left[f(X_k); X_k \in dx, T_1 > k\right] \\
&\quad \cdot \sum_{m=1}^{\infty} E\left[f(X_m); T_1 > m \mid X_0 = x, Y_0 = 0\right] \\
&\overset{(*)}{=} M \left[\int (f(x))^2 \pi(x) dx + 2 \sum_{m=1}^{\infty} \int\int \pi(x) f(x) Q^m(x, dy) f(y) dx \right] \\
&=: M\sigma_f^2,
\end{aligned}$$

missä yhtäsuuruus (*) on saatu käyttämällä yhtälöitä (5.3), (5.4), (5.10), (5.11) ja (5.17).

[3] $P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c) \leq P(A|B) + P(B^c)$

5.6 Markovin piilomalli

Usein todellisuudessa esiintyviä prosesseja ei voida mallintaa tavallisella Markovin ketjulla, sillä tilasiirtymät eivät välttämättä ole suoraan havaittavissa. Tällöin päätelmiä täytyy tehdä ns. tarkkailtavien tilojen perusteella. Markovin piilomalli (*Hidden Markov Model, HHM*) on tällaiseen tilanteeseen soveltuva äärellinen tilastollinen malli. Taustalla olevan Markovin ketjun tilat eivät ole suoraan havaittavissa, mutta ne synnyttävät havaintoja eli ns. emissioita, jotka muodostavat mallin toisen stokastisen prosessin. On huomattava, että sana "piilotettu" viittaa Markovin ketjun tiloihin, ei mallin parametrien estimoimiseen.

Esimerkki 5.23. (*Epärehellinen kasino*) Kuvitellaan tilanne, jossa epärehellinen kasino käyttää toisinaan tavallisen arpakuution sijasta epärehellistä noppaa. Tavallisen nopan todennäköisyys kullekin silmäluvulle on $1/6$, mutta epärehellinen noppa tuottaa silmäluvun 6 todennäköisyydellä $0,5$ ja muut silmäluvut todennäköisyydellä $0,1$. Oletetaan, että kasino vaihtaa epärehelliseen noppaan todennäköisyydellä $0,05$ ja takaisin tavalliseen noppaan todennäköisyydellä $0,1$. Tilasiirtymät noppien välillä muodostavat Markovin prosessin, jonka tilat eivät ole pelaajalle nähtävissä. Tilat synnyttävät kuitenkin emissioita, jotka tässä tapauksessa ovat nopanheiton tulokset kullakin pelikierroksella. Pelaajan näkökulmasta pelitapahtuma on Markovin piilomalli. [5, s.54, 1998.]

Määritelmä 5.24. Diskreetin Markovin piilomallin muodostaa kolmikko (V, S, λ), joka toteuttaa alla olevat viisi ehtoa. Parametrijoukko λ koostuu siirtymätodennäköisyysmatriiseista $(\mathbf{P}, \mathbf{B}, \pi)$.

1. Mallin tilajoukko S on N-alkioinen joukko
$$S = \{S_1, S_2, ..., S_N\}.$$
2. M alkioinen joukko symboleita muodostavat havaintoaakkoston $V = \{v_1, v_2, ..., v_M\}$. Jos havainnot ovat jatkuvia, niin M on ääretön.
3. Tilasiirtymätodennäköisyyksien $\mathbf{P} = \{\mathbb{P}(X_{n+1} = S_i \,|\, X_n = S_j) \,|\, S_i, S_j \in S\}$ tulee toteuttaa ehdot
$$\begin{aligned} \mathbb{P}(X_{n+1} = S_i \,|\, X_n = S_j) &\geq 0, \quad \text{kaikilla } 1 \leq i, j \leq N, n \geq 0, \\ \sum_{j=1}^{N} \mathbb{P}(X_{n+1} = S_i \,|\, X_n = S_j) &= 1, \quad \text{kaikilla } 1 \leq i \leq N, n \geq 0. \end{aligned}$$
4. Todennäköisyydet
$$b_j(k) = P\{Y_t = v_k \,|\, X_n = S_j\}, \quad 1 \leq j \leq N, \quad 1 \leq k \leq M,$$

missä X_n on prosessin tila ja Y_n emissio hetkellä n, muodostavat $N \times M$ matriisin $\mathbf{B}$.

5. Alkutodennäköisyysjakauma $\pi = \{\pi_i \,|\, S_i \in S\}$ määrittelee todennäköisyydet olla tilassa S_i hetkellä $n = 1$:

$$\pi_i = \mathbb{P}\{X_1 = S_i\}, \quad 1 \leq i \leq N.$$

5.7 Markovin ketjun Monte Carlo-menetelmä

Monte Carlo-menetelmät (*Monte Carlo, MC*) muodostavat joukon numeerisia algoritmeja, joiden toimintaperiaate perustuu satunnaisuuteen. Algoritmien avulla voidaan tuottaa simuloituja otoksia jostakin kiinnostavasta jakaumasta ja laskea simuloiduista tiloista suureita, joiden analyyttinen ratkaiseminen ei ole mahdollista. Kun jakaumalla on useampi kuin yksi ulottuvuus, satunnaislukujen sijaan poimitaan satunnaisvektoreita. Usein tämä on mahdollista toteuttaa vain nk. Markovin ketjun Monte Carlo-menetelmillä (*Markov Chain Monte Carlo, MCMC*).

Markovin ketjuja simuloivia algoritmeja kutsutaan Markovin ketjun Monte Carlo-menetelmiksi. Niiden avulla voidaan poimia satunnaisvektoreita jopa moniulotteisista jakaumista. MCMC-menetemien idea on löytää sellainen Markovin ketju, jonka tasapainojakauma on kiinnostuksen alla oleva todennäköisyysjakauma. Tällöin ketjun tiloja voidaan pitää riippuvina satunnaisvektoreina halutusta todennäköisyysjakaumasta.

5.7.1 Metropolis-Hastings-algoritmi

Tässä kappaleessa kuvataan, miten tunnetusta jakaumasta voidaan poimia empiirisiä otoksia Metropolis-Hastings-algoritmilla ja miksi algoritmi ylipäänsä on toimiva. Määrätköön parametrivektori θ havaittavissa olevan satunnaismuuttujan Y otantajakauman. Olkoon tällä jakaumalla jatkuva tiheysfunktio $p_Y(y|\theta)$. Jakauma voi myös olla diskreetti, jolloin puhutaan pistetodennäköisyysfunktiosta. Merkitään kiinnostuksen alla olevaa posteriorijakaumaa $\pi(\theta)$:lla, missä siis $\pi(\theta) = p(\theta|y)$.

Valitaan Markovin ketjun alkuarvoksi θ_0. Alkuarvon posterioritodennäköisyyden täytyy olla aidosti positiivinen eli $\pi(\theta_0) > 0$. Alkuarvon ei tarvitse olla jakautunut jakauman π mukaisesti [4, s.183]. Ehdokasjakaumasta $q(\cdot|\theta_{i-1})$ generoidaan satunnaisluku θ^*, jota sanotaan ehdokkaaksi ketjun seuraavalle tilalle. Ehdokas θ^* hyväksytään hyväksymissuhteella

$$\alpha(\theta^*; \theta_{i-1}) = \min\left\{1, \frac{\pi(\theta^*)q(\theta^*, \theta_{i-1})}{\pi(\theta_{i-1})q(\theta_{i-1}, \theta^*)}\right\}. \tag{5.31}$$

Voidaan kuvitella, että ehdokkaan hyväksymiseksi heitetään vinoutunutta noppaa, joka tuottaa kruunan todennäköisyydellä $\alpha(\theta^*;\theta_{i-1})$ ja klaavan todennäköisyydellä $1-\alpha(\theta^*;\theta_{i-1})$. Kruunan esiintyessä ehdokas θ^* hyväksytään ja asetetaan $\theta_i = \theta^*$. Klaavan esiintyessä ehdokas hylätään ja valitaan $\theta_i = \theta_{i-1}$. Tällöin ketjussa ei tapahdu hyppäystä. Iterointia toistetaan $H-1$ kertaa, jolloin saadaan ketju $(\theta_0, \theta_1, \ldots, \theta_H)$. Ketjulla on tasapainojakauma π, johon esitetään perustelut tuonnempana.

Bayesin-analyysin avulla posteriorijakauma π voidaan ilmaista muodossa

$$\pi(\theta) = p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y,\theta)d\theta} = \frac{f(\theta)}{C},$$

missä integraali C on vakio. Hyväksymissuhteessa (5.31) vakio C supistuu tällöin pois:

$$\alpha(\theta^*;\theta_{i-1}) = \min\left\{1, \frac{f(\theta^*)q(\theta^*,\theta_{i-1})}{f(\theta_{i-1})q(\theta_{i-1},\theta^*)}\right\}.$$

Metropolis-Hastings-algoritmin perusidea on juurikin se, ettei normalisointivakiota C tarvitse tietää.

Ketjun siirtymäydin on muotoa

$$K(\theta^*,\theta) := q(\theta^*|\theta)\alpha(\theta^*;\theta) + \delta_\theta(\theta^*)\left[1 - \int q(\theta^*|\theta)\alpha(\theta^*;\theta)d\theta^*\right], \tag{5.32}$$

missä ensimmäinen termi kuvaa todennäköisyyttä siirtyä tilasta θ tilaan θ^* ja jälkimmäinen termi todennäköisyyttä pysyä paikallaan tilassa θ ($\delta_\theta(\theta^*) = 1$, kun $\theta = \theta^*$, ja nolla muulloin).

Lause 5.25. *Markovin ketju, jolla on yhtälön (5.32) mukainen siirtymäydin $K(\theta^*,\theta)$, toteuttaa tasapainoehdon*

$$\pi(\theta^*)K(\theta^*,\theta) = \pi(\theta)K(\theta,\theta^*).$$

Todistus. Selvästi

$$\alpha(\theta^*;\theta)q(\theta|\theta^*)f(\theta^*) = \alpha(\theta;\theta^*)q(\theta^*|\theta)f(\theta)$$

ja

$$f(\theta^*)\delta_{\theta^*}(\theta)\left[1 - \int q(\theta|\theta^*)\alpha(\theta;\theta^*)d\theta\right] = f(\theta)\delta_\theta(\theta^*)\left[1 - \int q(\theta^*|\theta)\alpha(\theta^*;\theta)d\theta^*\right].$$

Laskemalla edelliset yhtälöt puolittain yhteen saadaan

$$\begin{aligned} f(\theta^*)K(\theta^*,\theta) &= f(\theta)K(\theta,\theta^*) \\ \Leftrightarrow \pi(\theta^*)K(\theta^*,\theta) &= \pi(\theta)K(\theta,\theta^*). \end{aligned}$$

□

Lause 5.26. *Edellä kuvatun Markovin ketjun tasapainojakauma on π eli*

$$\int \pi(\theta)K(\theta,\theta^*)d\theta = \pi(\theta^*).$$

Todistus.

$$\begin{aligned}\int \pi(\theta)K(\theta,\theta^*)d\theta &= \int \pi(\theta^*)K(\theta^*,\theta)d\theta \\ &= \pi(\theta^*)\underbrace{\int K(\theta^*,\theta)d\theta}_{=1} \\ &= \pi(\theta^*).\end{aligned}$$

□

Lauseen (5.26) mukaan todennäköisyys siirtyä tilaan θ^* on likimain $\pi(\theta^*)$ riippumatta siitä, mistä aloitusarvosta aloitetaan. Mitä pidemälle ketjussa edetään, sitä tarkemmin ketjun tilat ovat otoksia jakaumasta π. Tämä vaatii kuitenkin oletuksen, että ketju on pelkistymätön. Tämä tarkoittaa sitä, että todennäköisyys päästä äärellisellä määrällä siirtymiä jokaiseen parametrijoukon pisteeseen on positiivinen.

6 Markovin piilomalli lähijunaliikenteessä

6.1 Mallin parametrit

Aloitetaan tarkastelu tekemällä merkintöjä ja määrittelemällä mallin parametrit. Olkoon erään junalinjan pysähtymisasemat $0, 1, \ldots, T$. Olkoon tästä junalinjasta M riippumatonta ja samoinjakautunutta näytettä. Kaikilla $0 \leq a < b \leq T$ merkitään muuttujalla $X_a^b(i)$ lähdön i niiden matkustajien lukumäärää, jotka ovat nousseet kyytiin asemalta a ja jotka poistuvat asemalla b. Koska matkustajia ei ole merkitty, vektorit

$$X(i) = \left(X_a^b(i) : 0 \leq a < b \leq T\right), \quad i = 1, \ldots, M$$

eivät ole aineistosta havaittavissa.

Olkoon matkalle i

$$N_a(i) = \sum_{b=a+1}^{T} X_a^b(i), \tag{6.1}$$

asemalta a kyytiin nousseiden matkustajien kokonaislukumäärä ja

$$N^b(i) = \sum_{a=0}^{b-1} X_a^b(i). \tag{6.2}$$

asemalla b poistuneiden matkustajien kokonaislukumäärä. Havaittavissa oleva data koostuu mittauksista

$$N(i) = \Big(N_a(i), N^b(i) \,:\, 0 \le a < T, 0 < b \le T\Big), \quad i = 1, \ldots, M.$$

Käytetään lisäksi merkintää

$$Y_a^b(i) = N_a(i) - \sum_{a<k\le b} X_a^b(i)$$

kuvaamaan niiden matkustajien lukumäärää, jotka ovat nousseet kyytiin asemalta a, ja jotka ovat edelleen kyydissä junan poistuessa asemalta b.

Olkoon

$$X = \begin{bmatrix} \sum_{i=1}^{M} X_0^1(i) & \sum_{i=1}^{M} X_0^2(i) & \ldots & \sum_{i=1}^{M} X_0^T(i) \\ 0 & \sum_{i=1}^{M} X_1^2(i) & \ldots & \sum_{i=1}^{M} X_1^T(i) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sum_{i=1}^{M} X_{T-1}^T(i) \end{bmatrix}.$$

kaikista matkoista $i = 1, \ldots, M$ muodostuva $T \times T$ matriisi, jonka rivit ja sarakkeet summautuvat asemittain havaittaviksi kokonaisnousijoiksi ja -poistujiksi. Matriisi X on yläkolmiomatriisi, koska muuttujat X_a^b määritellään olemaan 0, kun $a \ge b$.

Yksittäisten matkojen pituuksien oletetaan olevan riippumattomia keskenään. Lisäksi samalta asemalta kyytiin nousseiden matkustajien tekemien matkojen pituudet oletetaan olevan identtisesti jakautuneet. Jakauma voi olla riippuvainen matkustajan lähtöasemasta. Näillä oletuksilla mallin parametrit ovat

$$\pi_a^b = P(X_a^b(i) = 1 \,|\, N_a = 1), \quad 0 \le a < b \le T,$$

mikä tarkoittaa asemalta a kyytiin nousseen matkustajan todennäköisyyttä poistua asemalla b, ja

$$\Lambda_a^b = P(X_a^b(i) = 1 \,|\, N_a = 1, Y_a^{b-1} = 1), \quad 0 \le a < b < T,$$

mikä tarkoittaa asemalta a kyytiin nousseen matkustajan todennäköisyyttä poistua asemalla b, kun tiedetään, että hän on edelleen kyydissä junan poistuessa asemalta $b-1$. Parametrit π_a^b ja Λ_a^b ovat ekvivalentit, sillä

$$\begin{aligned} \pi_a^b &= \Lambda_a^b \prod_{a<k<b} (1 - \Lambda_a^k) \\ \Lambda_a^b &= \pi_a^b / \sum_{b\le k\le T} \pi_a^k. \end{aligned}$$

6.2 Malli

Olkoon eräästä linjasta M riippumatonta ja samoinjakautunutta lähtöä, joissa kussakin kokonaisnousijat ja -poistujat ovat yhtäsuuret. Olkoon tutkittavan linjan pysähtymisasemat $0, 1, \ldots, T$. Tavoitteena on estimoida lähtöasemalle $a \in 0, 1, \ldots, T-1$ vektorin $(X_a^b : b = 1, 2, \ldots, T)$ ehdollinen todennäköisyys ehdolla nousijat N_a.

Jokaiselle matkalle i on olemassa lineaarinen systeemi

$$AX(i) = N(i), \tag{6.3}$$

jossa on $d = (T+1)T/2$ tuntematonta ja $2T$ yhtälöä. A on $2T \times d$ matriisi, jonka solut saavat arvoja joukosta $\{0, 1\}$ siten, että yhtälöt (6.1) ja (6.2) toteutuvat. Vektorille $X(i) \in \mathbb{N}^d$ voi löytyä useampi ratkaisu, kun lineaarinen systeemi on alimäärätty eli $T > 2$.

Metropolis-Hastings-algoritmin avulla pyrimme tutkimaan malliparametrin Λ ja matriisin X yhteisposteriorijakaumaa

$$p(\Lambda, X \,|\, \text{data}) \quad = \quad \frac{\pi(\Lambda) p_\Lambda(X \,|\, \text{data})}{C}. \tag{6.4}$$

Normalisointivakio C on integraalilauseke, jota ei tässä yhteydessä tarvitse tietää. Vaikka se osattaisiinkin laskea, se supistuu pois Metropolis-Hastings-algoritmissa käytettävää hyväksymissuhdetta laskettaessa.

Malliparametrien Λ_{ab} priorina käytetään jotakin beta-jakaumaa

$$\pi_{\alpha,\beta}(\Lambda_{ab}) = \frac{\Lambda_{ab}^{\alpha-1}(1-\Lambda_{ab})^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du},$$

missä $\alpha, \beta > 0$.

Olkoon $\Lambda_{ab} \in [0, 1]$ kiinnitetty. Tällöin

$$\mathbb{P}\left(\sum_{i=1}^{M} X_a^b(i) = k \,|\, \sum_{i=1}^{M} Y_a^b(i) = m\right) = \binom{m}{k} \Lambda_{ab}^k (1-\Lambda_{ab})^{m-k}$$

on binomijakautunut parametreilla m ja Λ_{ab}. Koska yksittäisten matkojen pituuksien oletetaan olevan riippumattomia, on todennäköisyys $p_\Lambda(X)$ tällöin binomitodennäköisyyksien tulo.

Jos oletetaan, että havaittu data on virheetöntä, lineaariselle systeemille (6.3) on aina olemassa vähintään yksi ratkaisu. Tämä ratkaisu on vektorin $X(i)$ "todellinen arvo". Jokaiselle lähdölle i etsitään yksi ratkaisu ja muodostetaan niistä matriisi X_0. Metropolis-algoritmin avulla lähdetään luomaan Markovin ketjua, jossa ketjun seuraava tila saadaan aina edellisen avulla. Ketjun aloitusarvoksi valitaan X_0.

Samplataan $\Lambda_0 = (\Lambda_{ab})_{a<b}$ jakaumasta

$$\begin{aligned}\mathbb{P}(\Lambda_{ab} \,|\, X_0) &\propto \frac{1}{\mathrm{Beta}(\alpha,\beta)}\Lambda_{ab}^{\alpha-1}(1-\Lambda_{ab})^{\beta-1}\binom{m}{k}\Lambda_{ab}^{k}(1-\Lambda_{ab})^{m-k} \\ &\propto \Lambda_{ab}^{k+\alpha-1}(1-\Lambda_{ab})^{m-k+\beta-1}\end{aligned}$$

kaikilla $0 \leq a < b < T$. Huomataan, että jakauma on normalisointivakiota vaille $\mathrm{Beta}(k+\alpha, m-k+\beta)$ jakauma.

Arvotaan mielivaltainen rivipari k, k', $k \neq k'$, sekä sarakepari l, l', $l \neq l'$. Valitut rivit ja sarakkeet risteävät neljässä eri kohdassa. Ehdotus $\widetilde{X}_K$ ketjun seuraavalle tilalle saadaan muokkaamalla alkuarvoa X_0 risteyskohdissa kuten

$$\widetilde{X}_K = X_0 + K \begin{pmatrix} + & - \\ - & + \end{pmatrix}, \quad K \in \mathbb{Z}.$$

Muunnos summaa ja vähentää satunnaisesti valitun kokonaisluvun K risteyskohdista kuvatulla tavalla. Tällä menettelytavalla rivien ja sarakkeiden summat pysyvät ennallaan. Mikäli muunnos tuottaa soluihin negatiivisen luvun tai matriisi ei ole enää yläkolmiomatriisi, se hylätään ja ehdotusta lähdetään etsimään uudestaan valitsemalla uudet rivi- ja sarakeparit.

Kun ei-negatiivinen ehdotus $\widetilde{X}_K$ on löydetty, tehdään Metropolis-Hastings askel joko hyväksymällä tai hylkäämällä ehdotus. Jos ehdotus hyväksytään, ketjun seuraavaksi tilaksi valitaan $X_1 = \widetilde{X}_K$. Mikäli ehdotus hylätään, ketjussa ei tapahdu hyppäystä, eli seuraavaksi tilaksi valitaan $X_1 = X_0$. Hyväksyminen tapahtuu todennäköisyydellä

$$A(\widetilde{X}\,;\, X_0) := \min\left\{1, \frac{p_{\Lambda_0}(\widetilde{X}_K)}{p_{\Lambda_0}(X_0)}\frac{J(\widetilde{X}_K \to X_0)}{J(X_0 \to \widetilde{X}_K)}\right\},$$

missä Λ_0 on kiinteä ja siirtymäydin J määrittää todennäköisyyden päästä muunnoksella tilasta toiseen.

Ketjun muodostamista jatketaan päivittämällä matriisia X ja malliparametria Λ vuorotellen; Λ_n samplataan jakaumasta

$$\Lambda_n \sim \mathbb{P}(\Lambda \,|\, X_n),$$

jonka jälkeen tehdään uusi Metropolis-Hastings askel $X_n \to X_{n+1}$. Kun algoritmi pyörii tarpeeksi pitkälle, Metropolis-Hastings ketjun viimeisen tilan jakauma lähestyy tasapainojakaumaa.

Myöhemmin ilmestyvässä *The local train problem*-artikkelissa (Gasbarra G. et al.) matkan pituuden mallintamista käsitellään perusteellisemmin. Tässä tutkielmassa esitetyssä

mallinnuksessa matkustajalaskentalaitteiden oletettiin laskevat oikein, mikä on liian epärealistinen oletus tehtäväksi. Artikkelissa mallia laajennetaan mallintamalla laskentalaitteiden tekemää virhettä. Tällöin aineistosta voidaan hyödyntää myös niitä lähtöjä, joissa kokonaisnousijat ja -poistujat eivät ole yhtäsuuret.

Artikkelissa tarkastellaan myös Markovin ketjun pelkistymättömyyttä. Jotta Metropolis-Hastings algoritmi toimisi luotettavasti, täytyy todennäköisyys päästä jokaiseen parametrijoukon alkioon äärellisellä määrällä tilasiirtymiä olla positiivinen. Valitulla Metropolis-Hastings ehdotusjakaumalla on merkittävä vaikutus siihen, kuinka sujuvasti ketju pääsee liikkumaan parametriavaruudessa. Ehdotus ei saa olla liian lähellä eikä liian kaukana edellisestä arvosta. Edellisessä ketju liikkuu, mutta liian pienin askelin, kun taas jälkimmäisessä ketju voi pysyä pitkiäkin aikoja paikallaan.

Markov hypoteesin mukaan historialla ei ole merkitystä ketjun tuleviin tapahtumiin, mikäli ketjun nykytila on tiedossa. Tässä sovelluskohteessa se tarkoittasi sitä, että matkustajan lähtöasemalla ei ole väliä arvioitaessa tulevilla asemilla tapahtuvia poistujamääriä. Tätä hypoteesin testausta sekä muita tuloksia esitellään myös artikkelissa.

7 Lopuksi

Tämän tutkielman tarkoituksena oli kehittää tilastollinen menetelmä kompensoimaan lähijunaliikenteen matkustajamäärätutkimuksessa syntyvää vastauskatoa. Matkustajalaskentalaitteiden keräämä aineisto on vinoutunut ja sisältää erävastauskatoa, mikä asetti haasteita menetelmän kehittämiseen. Lisäksi tuli huomioida se, että tulosaineisto taipuu vähintään vastaavanlaiseen raportointiin kuin mitä manuaalilaskennoista on tehty.

Projektin aikaraja oli tammikuu 2013, mistä lähtien laskentalaitteisiin perustuva tilastointi oli määrä aloittaa. Matkustajamäärätutkimuksen prosessi valmistui raportointisovellusta lukuun ottamatta aikataulussa. Tämän tutkielman viimeistelyvaiheessa raportointisovellus oli juuri siirtymässä testausvaiheesta tuotantoon.

Lähitulevaisuudessa IVU-järjestelmän ja matkustajamäärätutkimuksen välisen rajapinnan määrittäminen tulee luultavasti ajankohtaiseksi. IVU-järjestelmästä voitaisiin saada junayksiköiden järjestys kokoonpanossa, mikä helpottaisi erityisesti erävastauskadon käsittelyä. Lisäksi sieltä saataisiin toteutunut aikataulu ja kalusto sekä peruuntuneet lähdöt, joiden avulla voitaisiin ehkäistä kehikkovirheitä.

Matkan pituuksien mallintaminen MCMC-menetelmillä on uudenlainen tapa analysoida matkustajakäyttäytymistä hyödyntämällä ainoastaan laskentalaitteiden tuottamaa dataa. Malli on herättänyt paljon kiinnostusta kansainvälisestikin. Aikaisintaan vuonna 2016

voimaan astuvan uuden taksa- ja lippujärjestelmän arvolipun hinnoitteluperiaate tulee perustumaan kuljetun matkan pituuteen. Siinä arvolipun käyttäjä tekee sisään-leimauksen kulkuneuvoon noustessaan ja ulos-leimauksen aina kulkuneuvosta poistuessaan. [27, s.18-19, 2009.] Tulevaisuus näyttää, onko matkan pituuksien mallintaminen laskentalaitteiden tuottaman aineiston perusteella enää aiheellista.

Lähdeluettelo

[1] Acuña E. & Rodriguez C. 2004. The treatment of missing values and its effect in the classifier accuracy. Classification, Clustering and Data Mining Applications, 639-648.

[2] DavisWeb 2012. Dilax Intelcom GmbH.

[3] DavisWebin Help-valikko 2012. Dilax Intelcom GmbH.

[4] Diaconis P. 2008. The Markov Chain Monte Carlo Revolution. Bulletin of the American Mathematics Society 2, 179-205.

[5] Durbin R. & Eddy S. & Krogh A. & Mitchison G. 1998. Probabilistic Models of Proteins and Nucleic Acids. Gambridge: Gambridge University Press.

[6] Engel M. 2012. Sateenvarjomäärittelyn 2. Workshop 23.2.2012.

[7] Engel M. 2013. Head of Division, Dilax GmbH. DILAX Manual counting - how to -. Vastaanottaja marting.engel@dilax.com. Lähetetty 25.02.2013.

[8] Gasbarra D. & Bergandi G. & Nikula N. n.d. The local train problem.

[9] Helsingin seudun liikenteen www-sivut 2012. http://www.hsl.fi.

[10] Hurmerinta J. 2012. Suunnittelupäällikkö, VR-Yhtymä Oy. Kalustomuutokset. Vastaanottaja joona.hurmerinta@vr.fi. Lähetetty 21.11.1012.

[11] Illenberger J. & Flötteröd G. & Nagel K. 2009. An approach to correct biases induced by snowball sampling. http://svn.vsp.tu-berlin.de/repos/public-svn/publications/vspwp/2008/08-16/snowball.pdf.

[12] Into L. 2010. Joukkoliikenteen matkustajalaskentajärjestelmät. Diplomityö. Aalto-yliopisto.

[13] Koralov L. B. & Sinai Y.G. 2007. Theory of probability and random processes. Berlin: Springer.

[14] Laaksonen S. 2010. Surveymetodiikka. Frederiksberg: Ventus Publishing ApS.

[15] Lehtonen R. ja Pahkinen E. 2004. Practical Methods for Design and Analysis of Complex Surveys. New York: John Wiley & Sons.

[16] Liikennevirasto 2012. Rautatieliikenteen täsmällisyys 2011. Verkkojulkaisu: www.liikennevirasto.fi.

[17] Lindvall A. 2012. Liikennesuunnittelija, VR-Yhtymä Oy. Vanhat manuaalilaskennat. Vastaanottaja antti.lindvall@vr.fi. Lähetetty 26.02.2012.

[18] Nummelin E. 2002. MC's for MCMC'ists. International Statistical Review 2, 215-240.

[19] Pardoux E. 2008. Markov Processes and Applications. Chichester: John Wiley & Sons, Inc.

[20] Parzen E. 1999. Stochastic Processes. Society for Industrial and Applied Matehematics.

[21] Pfefferman D. & Rao C.R. 2009. Handbook of Statistics 29: Sample surveys. North-Holland.

[22] Sottinen T. 2006. Todennäköisyysteoria. Kurssin luentomateriaali. http://intmath.org/home/tsottine/?download=tn.pdf.

[23] Tilastokeskus 2007. Laatua tilastoissa, 2. uudistettu painos. Helsinki: Yliopistopaino. http://www.stat.fi/meta/qg_2ed.pdf.

[24] de Waal T. & Pannekoek J. & Scholtus S. 2011. Handbook of statistical data editing and imputation. New Jersey: John Wiley & Sons, Inc.

[25] Westpahl S. 2012. Customer Service, Dilax GmbH. Palaveri Berliinissä 22.8.2012.

[26] Westpahl S. 2012. Customer Service, Dilax GmbH. Non-matched trips. Vastaanottaja sven.westphal@dilax.com. Lähetetty 22.10.2012 klo. 09:39.

[27] YTV 2009. Pääkaupunkiseudun joukkoliikenteen taksa- ja lippujärjestelmän 2014 alustava kuvaus. Helsinki: Valopaino.